

LINGUISTICA  
COMPUTAZIONALE

VOLUME IX · X

---

CURRENT ISSUES  
IN COMPUTATIONAL  
LINGUISTICS:  
IN HONOUR OF DON WALKER

EDITED BY:  
ANTONIO ZAMPOLLI, NICOLETTA CALZOLARI,  
MARTHA PALMER

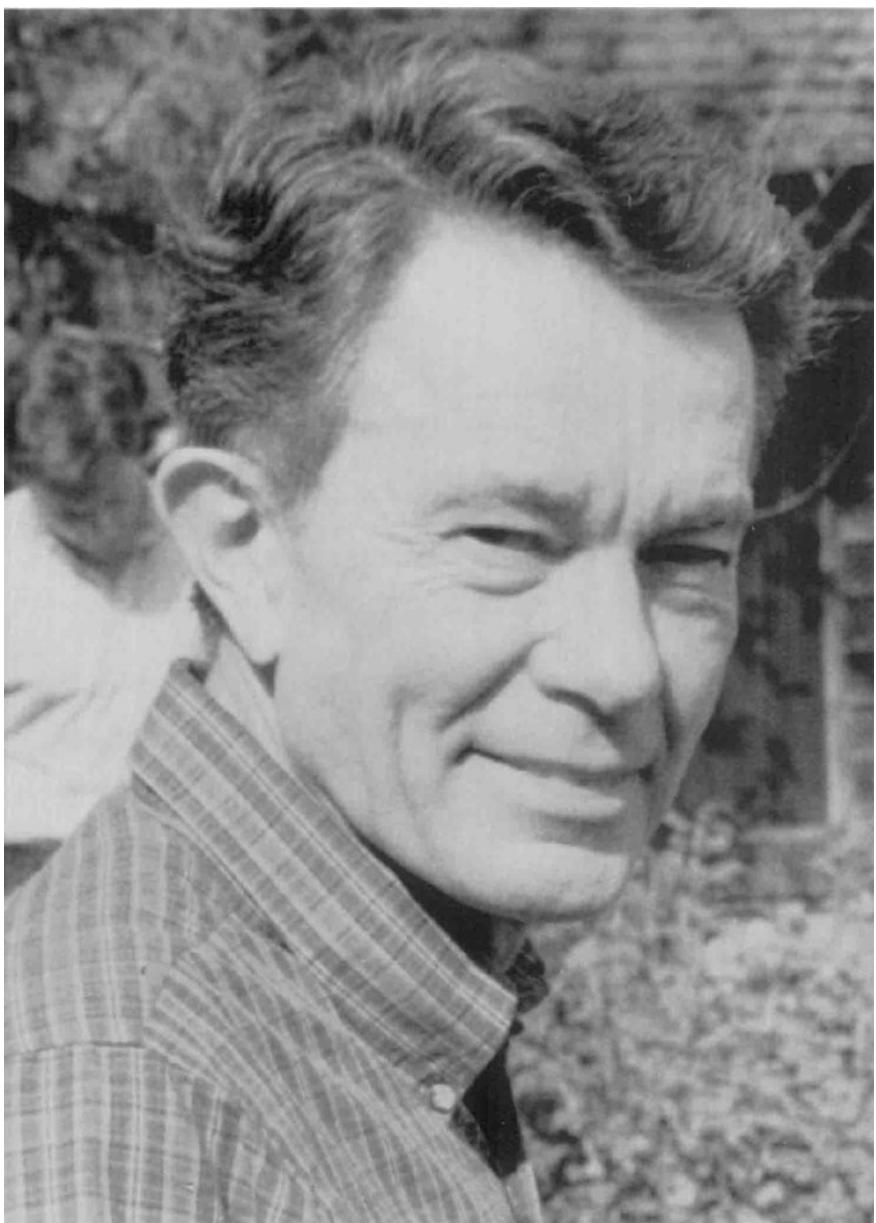


GIARDINI EDITORI  
E STAMPATORI  
IN PISA



SPRINGER-SCIENCE+  
BUSINESS MEDIA, B.V.

**LINGUISTICA  
COMPUTAZIONALE**



LINGUISTICA  
COMPUTAZIONALE

VOLUME IX · X

---

CURRENT ISSUES  
IN COMPUTATIONAL  
LINGUISTICS:  
IN HONOUR OF DON WALKER

EDITED BY:  
ANTONIO ZAMPOLLI, NICOLETTA CALZOLARI,  
MARTHA PALMER



Springer-Science+Business Media, B.V.

*Linguistica computazionale* is published two 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.

---

This book is co-published by *Giardini editori e stampatori* in Pisa and *Kluwer Academic Publishers*.  
*Kluwer Academic Publishers* incorporates the publishing programmes of D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

© Springer Science+Business Media Dordrecht 1994

Originally published by *Giardini editori e stampatori* in Pisa in 1994

Softcover reprint of the hardcover 1st edition 1994

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, electronic, mechanical, photocopying, recording, or otherwise, without written permission.

Library of Congress Cataloging-in-Production Data available

ISBN 978-0-7923-2998-5      ISBN 978-0-585-35958-8 (eBook)

DOI 10.1007/978-0-585-35958-8

---

Royalties will be made available by *Kluwer Academic Publishers* to the Association of Computational Linguistics Don and Betty Walker International Student Fund.

---

## CONTENTS

Preface	IX
Introduction	XVII
Donald Walker: A Remembrance	XXI
 <b>SECTION 1 - THE TASK OF NATURAL LANGUAGE PROCESSING</b>	
K. SPARCK JONES, Natural Language Processing: A Historical Review	3
J. ROBINSON, On Getting a Computer to Listen	17
B. GROSZ, Utterance and Objective: Issues in Natural Language Communication	21
M. KING, On the Proper Place of Semantics in Machine Translation	41
G. G. HENDRIX, E. D. SACERDOTI, D. SAGALOWICZ, J. SLOCUM, Developing a Natural Language Interface to Complex Data	59
K. KUKICH, K. McKEOWN, J. SHAW, J. ROBIN, J. LIM, N. MORGAN, J. PHILLIPS, User-Needs Analysis and Design Methodology for an Automated Document Generator	109
 <b>SECTION 2 - BUILDING COMPUTATIONAL LEXICONS</b>	
B. BOGURAEV, Machine-Readable Dictionaries and Computational Linguistics Research	119
R. A. AMSLER, Research Toward the Development of a Lexical Knowledge Base for Natural Language Processing	155
R. J. BYRD, Discovering Relationships among Word Senses	177
E. HAJIČOVÁ, A. ROSEN, Machine Readable Dictionary as a Source of Grammatical Information	191
S. PIN-NGERN CONLON, J. DARDAIN, A. D'SOUZA, M. EVENS, S. HAYNES, J. S. KIM, R. STRUTZ, The IIT Lexical Database: Dream and Reality	201
J. L. KLAVANS, Visions of the Digital Library: Views on Using Computational Linguistics and Semantic Nets in Information Retrieval	227
B. T. ATKINS, J. KEGL, B. LEVIN, Anatomy of a Verb Entry: from Linguistic Theory to Lexicographic Practice	237
N. CALZOLARI, Issues for Lexicon Building	267
N. IDE, J. LE MAITRE, J. VÉRONIS, Outline of a Model for Lexical Databases	283
L. LEVIN, S. NIRENBURG, Construction-Based MT Lexicons	321

P. SGALL, Dependency-Based Grammatical Information in the Lexicon	339
H. SCHNELLE, Semantics in the Brain's Lexicon - Some preliminary Remarks on its Epistemology	345
<b>SECTION 3 - THE ACQUISITION AND USE OF LARGE CORPORA</b>	
D. E. WALKER, The Ecology of Language	359
D. BIBER, Representativeness in Corpus Design	377
C. M. SPERBERG-MCQUEEN, The Text Encoding Initiative	409
W. A. GALE, K. W. CHURCH, D. YAROWSKY, Discrimination Decisions for 100,000-Dimensional Spaces	429
S. ARMSTRONG-WARWICK, Acquisition and Exploitation of Textual Resources for NLP	451
S. HOCKEY, The Center for Electronic Texts in the Humanities	467
N. J. BELKIN, Design Principles for Electronic Textual Resources: Investigating Users and Uses of Scholarly Information	479
<b>SECTION 4 - TOPICS, METHODS AND FORMALISMS IN SYNTAX, SEMANTICS AND PRAGMATICS</b>	
A. K. JOSHI, Some Recent Trends In Natural Language Processing	491
J. R. HOBBS, J. BEAR, Two Principles of Parse Preference	503
M. NAGAO, Varieties of Heuristics in Sentence Parsing	513
R. JOHNSON, M. ROSNER, UD, yet another Unification Device	525
J. FRIEDMAN, D. B. MORAN, D. S. WARREN, Evaluating English Sentences in a Logical Model	535
M. S. PALMER, D. A. DAHL, R. J. SCHIFFMAN, L. HIRSCHMAN, M. LINEBARGER, J. DOWDING, Recovering Implicit Information	553
C. L. PARIS, V. O. MITTAL, Flexible Generation: Taking the User into Account	569
Y. WILKS, Stone Soup and the French Room	585

## Preface

*With this volume in honour of Don Walker, Linguistica Computazionale continues the series of special issues dedicated to outstanding personalities who have made a significant contribution to the progress of our discipline and maintained a special collaborative relationship with our Institute in Pisa. I take the liberty of quoting in this preface some of the initiatives Pisa and Don Walker have jointly promoted and developed during our collaboration, because I think that they might serve to illustrate some outstanding features of Don's personality, in particular his capacity for identifying areas of potential convergence among the different scientific communities within our field and establishing concrete forms of cooperation. These initiatives also testify to his continuous and untiring work, dedicated to putting people into contact and opening up communication between them, collecting and disseminating information, knowledge and resources, and creating shareable basic infrastructures needed for progress in our field.*

*Our collaboration began within the Linguistics in Documentation group of the FID and continued in the framework of the ICCL (International Committee for Computational Linguistics). In 1982 this collaboration was strengthened when, at COLING in Prague, I was invited by Don to join him in the organization of a series of workshops with participants of the various communities interested in the study, development, and use of computational lexica. I was unable to participate in the first workshop organized by Don at SRI in 1983, because I was involved in a CETIL meeting<sup>1</sup> at the same time in Luxembourg, where the suggestion of holding a second workshop in Europe was accepted. So, Don, Nicoletta Calzolari and I, together with Loll Rolling (CETIL promoter) and Juan Sager (CETIL President), organized the workshop "On Automating the Lexicon" in 1986 in Grosseto, sponsored by the CEC, the Council of Europe, ACL, AILA, ALLC and EURALEX. The objective of this workshop was to survey research efforts, current practices, and potential developments in work on the lexicon, machine-readable dictionaries, and lexical knowledge bases, with special consideration of the problems created by working in a multilingual environment. The brief was to recommend directions for future activities. The participants were chosen to bring together, for the first time, representatives of all those working on the lexicon: lexicologists, grammarians, semanticists, lexicographers, computational linguists, artificial intelligence specialists, cognitive scientists, publishers, lexical software producers, translators, terminologists, and representatives of funding agencies and of professional associations. The final recommendations, transmitted to the CEC*

---

<sup>1</sup>CETIL was a CEC Committee of Experts in Language and Information.

*and widely distributed, could be summarized as follows:* <sup>2</sup>

- *To establish procedures for creating multifunctional databases from the information contained both implicitly and explicitly in those traditional dictionaries that exist in machine-readable form.*
- *To develop computational tools for more efficient handling of lexical and lexicographical data, and to provide 'workstation' environments within which these tools may be used by lexicologists and lexicographers.*
- *To explore the possibility of creating multifunctional lexical databases capable of general use, despite divergences of linguistic theories and differences in computational and applicational frameworks.*
- *To study the possibility of linking lexical databases and large text files, in both monolingual and multilingual contexts, in order to determine the most effective ways of exploiting the relationships among the various lexical elements.*

*Don dedicated an unbelievable amount of time, effort, and care in preparation of this workshop. Don, Nicoletta and I were in touch nearly every day for more than a year. I remember this, with joy, as one of the happiest and most fruitful periods of my working life. During the organization of the workshop, and as we went along establishing relationships with the participants invited to it, we had the feeling that a new research paradigm was emerging more and more clearly every day. In Pisa, we found here a confirmation of our work on texts, reusable lexica, and related tools, despite the almost complete lack of interest in the major trends in contemporary computational linguistics. Don found continual confirmations for his intuition of the multi-disciplinary centrality of the lexicon, and for the necessity of a vast organization to provide the various communities, represented in the workshop, with basic linguistic resources.*

*The Grosseto Workshop is now recognized as marking the starting-point of a new phase in the field of computational linguistics. This phase is characterized by an increasing number of fresh initiatives, particularly at the international level, whose aim is to further the development of the scientific, technical, and organizational conditions conducive to the creation of large multifunctional linguistic resources, such as written and spoken corpora, lexicons and grammars. An updated version of the proceedings of the Grosseto Workshop, edited by Don, Nicoletta and me, will appear soon (Walker et al, 1994). Don put a lot of effort into collecting and updating the articles, and we are extremely grateful that he was able to finish this task.*

*Don was particularly active in promoting actions whose purpose was to realise the strategic vision that emerged during the workshop. He organized two follow-up workshops, one in New York in 1986, the other, in 1987, in Stanford, in the framework of the Linguistic Summer School of the LSA.*

---

<sup>2</sup>See Zampolli, 1987, for the complete text of the recommendations.

*He also organized a panel on linguistic resources at the 1988 Pisa Summer School on Computational Lexicography and Lexicology, explicitly designed to contribute to the goals identified during the Grosseto Workshop. Unfortunately, the day before the panel was due to meet, Don notified me that he was not able to come, because his illness had just been discovered and he was about to have his first operation. I still remember that phone call with deep emotion. Don encouraged me to carry on with the work regardless. The way in which, during the five years of his illness, he was able to continue working with energy and enthusiasm, guiding and encouraging us in the various initiatives we were promoting together, and always planning new activities, was for me a miraculous example of what devotion to one's discipline, community and ideals can really achieve.*

*The day after the Grosseto Workshop, I set up a group (Hans Uszkoreit, Nicoletta Calzolari, Bob Ingris, Bran Boguraev) to explore the feasibility of constructing large-scale linguistic resources, explicitly designed to be multifunctional, i.e. capable of serving, through appropriate interfaces, a wide variety of present and future research and applications. A crucial and controversial problem was to what extent it was possible, as well as desirable, to make linguistic resources, at least within certain limits, "polytheoretical", i.e. usable in different linguistic frameworks. Don immediately intervened and secured for the group, initially supported by our Institute, the help of the ACL. He was enthusiastic about the goal of this group, (the so-called "Pisa - Group"), which was extended to include outstanding representatives of some major linguistic schools, and (I quote Don's words) "investigated in detail the possibility of a polytheoretical representation of the lexical information needed by parsers and generators, such as the major syntactic categories, subcategorization and complementation. The common representation sought was one that could be used in any of the following theoretical frameworks: government and binding grammar, generalized phrase structure grammar, lexical functional grammar, relational grammar, systemic grammar, dependency unification grammar and categorial grammar" (Walker et al, 1987).*

*After the Grosseto Workshop, we wrote a report together for the CEC DG-XIII, suggesting a large two-phase programme: a one-year phase, to define the methods and the common specifications for a coordinated set of lexical data bases and corpora for the European languages, and a second three-year phase for their actual construction (Zampolli, Walker, 1987). This programme, it seems, is finally being realized now, even if organizational and financial problems have made its progress slower than thought.*

*The first step taken by the CEC consisted in a feasibility study: the ET-7 Project, building on the encouraging results of the "Pisa-Group" work. This project was launched by the CEC with the aim of recommending a methodology for the concrete construction of shareable lexical resources. Since different theories use different descriptive devices to describe the same linguistic phenomena and yield different generalizations and conclusions, ET-7 proposed the use of the observable differences between linguistic phenomena as a platform for the exchange of data. In particular, the study has assessed the feasibility of some basic standards for the description of lexical items at the level of orthography, phonology/phonetics, morphology, collocation, syntax, semantics and pragmatics.*

(Heid, McNaught 1991)

The second major step is now underway, and is represented by the establishment of projects focussing on the notion of standards. Two examples in the field of lexical data are the CEC ESPRIT project MULTILEX, whose objective was to devise a model for multilingual lexicons (Khatchadourian and Modiano, 1994), and the EUREKA project GENELEX, which concentrates on a model for monolingual generic lexicons (Antoni-Lay, Francopoulo and Zaysser, 1993). In the area of textual corpora the CEC sponsored the NERC study aiming at defining the scientific, technical and organizational conditions for the creation of a Network of European Reference Corpora, and at exploring the feasibility of reaching a consensus on agreed standards for various aspects of corpora building and analysis (see NERC Final Report, 1993). The European Speech community had independently organized an outstanding standardization activity, coordinated through the ESPRIT Project SAM (Fourcin, Gibbon, 1993). Feeling it necessary to coordinate their activities, the representatives of these various standardization projects formed an initial, preparatory group. Enlarging this group, the CEC established the EAGLES project in the framework of the LRE programme. This project aims to provide guidelines and de facto standards, based on the consensus of the major European projects, for the following areas: corpora, lexica, formalism, assessment and evaluation and speech data. (*Elsnews Bulletin*, 2(1), 1993). The project also encompasses an international dimension, influenced by Don's suggestions, which includes: the support of the European participation in the TEI; the preparation of a survey of the state-of-the-art in Natural Language and Speech Processing, jointly sponsored by the NSF and the CEC; the preparation of a Multilingual Corpus (MLCC) intended to support Cooperation with similar ARPA sponsored initiatives; and the exploration of possible strategies for international cooperation and coordination in the field of linguistic resources.

We are now on the verge of the third step needed to complete the programme suggested in our 1987 report to the CEC: to define and provide a common set of basic multifunctional reusable linguistic resources for all the European languages, available in the public domain, and to create an infrastructure for the collection, dissemination and management of new and existing resources. The second task is taken care of, at the experimental level, by the new LRE-RELATOR project (*Elsnews Bulletin*, 2(5), 1993). The MLAP call issued recently by the CEC DG-XIII, gives us the hope that the first task will be finally undertaken within the 4th Framework Research Programme of the CEC.

Given his interest for making linguistic resources shareable, and people and disciplines cooperate, the deep involvement of Don in standardization efforts is not surprising. His participation, from the very beginning, in the process of setting up the Text Encoding Initiative, is yet another testimony to his capacity to overcome (often artificial) disciplinary boundaries and of his determination in realizing his strategic vision of the future. Nancy Ide, President of the ACH, in November 1987 held a brainstorming workshop of representatives of the ACH and ALLC at Vassar to explore the desirability and feasibility of standards for encoding and exchanging texts in the field of humanities. Don, informed of this event, contacted Nancy and was also invited together with Bob Amsler. I remember my

*emotion on hearing Don (in the evening I had gathered Don round the fire together with Susan Hockey and Nancy Ide, to discuss the possibilities of cooperation of the three Associations) announcing that he would ensure the support of the ACL to a joint initiative with the ACH and ALLC. Those of us old enough, like me, to have lived through the entire history of linguistic data processing, will probably understand my reaction. In the '50s and early '60s, activities such as machine translation, document retrieval, lexical text analysis for humanities, statistical linguistics, quantitative stylistics, which later on were separated by scientific and organizational barriers, were linked by frequent contacts and recognized each other as "poles" of a disciplinary continuum, linked together by an interest in computational processing of large "real" texts and language data. In the second half of the '60s, computational linguistics, in an effort to define its disciplinary identity, and under the influence of the generative paradigm in linguistics and of the embryonic AI, became progressively detached from the interest and work on "real" language data. The occasions for contact with the community of humanities computing became increasingly rare, despite the efforts of some institutes, which refused this separation, and continued to consider the computational processing of real language data as a unifying goal to which the various disciplines could contribute relevant know-how, methods and tools.*

*Reversing this trend, Don went on to play a central role within the TEI, where he contributed his capacities to mediate, to organize and balance our efforts, and to envision the future. Despite our tremendous loss, we are determined to realize his vision of the TEI as a bridge between standardization activities in various continents. He also recognized the areas of potential synergies between computational linguistics and literary and humanistic computing, and identified the benefits potentially deriving from mutual cooperation. He promoted, in particular through the friendships established in the framework of TEI with Susan Hockey (chairman of the ALLC), various events jointly sponsored by the ACH, ACL, ALLC, and above all, he actively contributed, from the initial steps to the establishment of the CETH (Center for Electronic Texts in Humanities).*

*Don's interest in linguistic resources (conceived as a fundamental infrastructure for natural language and speech processing) and standards (conceived as a means for allowing exchanges and cumulative efforts) naturally converged with his vision of "Language, information and knowledge....combined in human efforts after communication"<sup>3</sup> and with the profound humanism that pervaded all his activities. At the basis of his capacity to open up lines of communication between peoples and communities and to propose new collaborative initiatives, of his will to create the best working conditions for everyone, of his continuous and untiring work in the promotion of scientific development in our sector, considered by him to be inseparable from the development of people, lie, I believe, not only his desire to open new perspectives maximizing synergies and interdisciplinary fertilization, but also and above all his genuine, profound interest in people. In the various activities of his long career, he was always inspired by a sense of help, and by a deeply democratic and humanistic conception of research and its management.*

*I would like to relate a personal anecdote which revealed to me Don's strong*

---

<sup>3</sup>See Don's article in this volume.

*values. One evening, on the occasion of a reunion of the "Pisa Group", I invited Don to my house, with Annie Zaenen, to listen to some music, and at a certain moment I put on the records of "The Weavers Reunion at Carnegie Hall" of 1953 and 1963. I was surprised to see that Don became more and more moved as each song played, and I asked him why. He told me that the Weavers remained, at the beginning of his scientific career, during the worst period of the cold war, one of the most visible symbols of democracy and freedom of culture in his country, the basic values that had inspired his life.*

*To close, I would like to thank, in the name of our journal *Linguistica Computazionale*, all those who have contributed to the realization of this issue, moved by admiration, gratitude and love for Don and his work. The Scientific Council of the ILC, in its role as editorial board of the journal, has approved and sustained my proposal to dedicate this number to Don Walker. Nicoletta Calzolari, Susan Hockey, James Pustejovsky and Susan Armstrong have signed with me the invitation to the contributors. As co-editor of this volume, Martha Palmer expended substantial scientific and organizational effort in collecting and preparing these articles, for which our heartfelt thanks. We are grateful as well to Carolyn Elken, Tarina Ayazi and Antonietta Spanu for their precious assistance.*

*Antonio Zampolli*

*I only wish to express my deep gratitude to Don, with very simple words, and of my many different memories of him just touch on one or two of those which perhaps mean most to me. I strongly feel the need to thank Don for what I received from him as a friend, as a fellow linguist, and as a person.*

*As a friend he treated me as I later realised he treated others whom he trusted: he helped me, he encouraged me, he gave me the right opportunities in the right moments, and I always knew that I could rely on him, at any moment. And this is a very rare gift indeed. We worked together in a wonderful way, as only friends can do.*

*As a linguist, we all know what a great contribution he made to our field and in how many different ways, both big and small: moving things forward, putting people together, encouraging newcomers, understanding in advance what would be important in the coming years, and acting concretely to make things happen. Among his many talents, he was a great builder.*

*As a person, I really must express my immense admiration for the way in which, once he knew of his illness and right up to the last days, he not only engaged in his own very private but constant battle but was constantly and courageously able to talk about the future with others, to make plans, to act towards the achievement of results that he knew that he personally would never see. And always with a smile and a kind light in his eyes. This for me was the clearest sign of his greatness. This memory will remain with me forever.*

*Nicoletta Calzolari*

## References

- [1] Antoni-Lay, M.H., G. Francopoulo, L. Zaysser, "A Generic Model for Reusable Lexicons: The Genelex Project", *Literary and Linguistic Computing*, 8 (4), Oxford University Press, 1993.
- [2] "RELATing to Resources", *Elsnews Bulletin*, 2 (5), 1993.
- [3] "EAGLES Update", *Elsnews Bulletin*, 2 (1), 1993.
- [4] Fourcin, A., D. Gibbon, "Spoken Language Assessment in the European Context", *Literary and Linguistic Computing*, 8 (4), Oxford University Press, 1993.
- [5] U. Heid, J. McNaught (eds.), "Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications", *Eurotra-7 Final Report*, Stuttgart, 1991.
- [6] Khatchadourian, H., N. Modiano, "Use and Importance of Standard in Electronic Dictionaries: The Compilation Approach for Lexical Resources", *Literary and Linguistic Computing*, 8 (4), Oxford University Press, 1993.
- [7] LRE, "Technical Report Document", Unpublished document available from R. Cencioni, CEE- DGXIII, Luxembourg.
- [8] NERC Final Report, Istituto di Linguistica Computazionale, Pisa, 1993.
- [9] Walker, D., A. Zampolli, N. Calzolari, "Toward a polytheoretical Lexical Database", Istituto di Linguistica Computazionale, Pisa, 1987.
- [10] Walker, D., A. Zampolli, N. Calzolari (eds.), *On Automating the Lexicon: Research and Practice in a Multilingual Environment*, Proceedings of a Workshop held in Grosseto, Oxford University Press, Oxford, 1994.
- [11] Zampolli A., Walker D., "Multilingual Lexicography and Lexicology; New Directions", Unpublished paper presented to CGXII-TSC of the CEC.
- [12] Zampolli A., "Perspectives for an Italian Multifunctional Lexical Database", *Studies in honour of Roberto Busa S.J.*, A. Zampolli, A. Cappelli, C. Peters (eds.), *Linguistica Computazionale*, Pisa, 6, 1987, 301-341.
- [13] Zampolli A., "Technology and Linguistic Resources", in M. Katzen (ed.), *Scholarship and Technology in the Humanities*, British Library Research, London, 1991.

# Introduction

This collection of papers is dedicated to Don Walker, a kind, gentle soul, and a man of intelligence and vision whom we have been privileged to know and to work with. The technical content of these papers clearly reflects his guidance and direction: his understanding of the relevance of theories from related disciplines such as linguistics, psychology, and philosophy; and his dream of the future availability of electronic documents and the potential of this as a resource for corpus-based analysis. This led to one of his most interdisciplinary roles, as a primary mover in the push to standardize the acquisition and tagging of electronic text, an endeavor that brought together people from fields as distinct as publishing, the humanities and computer science. His wide-spread interests are mirrored in the rich and diverse collection of papers in this volume which portray many different approaches and many different styles. These papers co-exist in harmony here, united by the goal of paying tribute to Don as well as actualizing part of his dream.

Anyone perusing this book will want to start with Karen Sparck Jones's entertaining and insightful history of natural language processing. Not only does she aptly sum up the driving forces behind particular developments in the field, but her history provides an appropriate context in which to appreciate the rest of the book. Karen sees the initial impetus for the field beginning with the push to perform automatic machine translation. It has been followed by three separate phases of development, each of which can be characterized by a distinct methodological approach, as researchers have explored the incorporation of techniques and theories from other disciplines. The first of these was focused on building knowledge based natural language systems, the second one, the grammatico-logical phase, was aimed at identifying an appropriate granularity of logical formalisms for representing syntactic and semantic information, and finally, our current phase, based on making the best possible use of the vast quantities of text that are now available electronically. In spite of the differences, she sees common themes running through all four phases, such as the balancing of generalizable principled methods with the "ad-hoc particularization necessary for specific applications," "the relative emphasis to be placed on syntax and on semantics" and an attempt to measure "the actual value of the results, especially when balanced against pre- or post- editing requirements." She also sees a certain inevitability in the field's return to machine translation as one of its preeminent application areas, and the subsequent interest in cross-linguistic studies.

Our first section, *The Task of Natural Language Processing*, begins with Karen's paper, which is followed by three classic articles highlighting some of the most intractable problems in the field, the solutions to which still lie tantalizingly out of reach, and which transcend any particular methodological approach: Jane Robinson describes the complexities inherent in understanding spoken language; Barbara Grosz discusses the importance of shared models for effective communication and the distinction between an utterance and the objective of its speaker; and Maghi King explores linguistic evidence

supporting the presence or absence of semantic universals. The difficulties inherent in the task of natural language processing have not daunted many intrepid system builders, and this section concludes with two different real-world applications, both of which have had to make reference to the balancing of the principled methods exemplified by the preceding papers and "ad-hoc particularization." The first application is a classic database question answering system developed by Gary Hendrix, et al., in the late 70's under Don's leadership. It is followed by a current, state of the art, documentation generation system from Don's more recent colleagues at Bellcore, Karen Kukich, et al.

The sections that follow focus more on the particular modules and components that these large-scale systems are comprised of, with special emphasis being placed on lexical issues and deriving information automatically from dictionaries and corpora. Our second section, *Building Computational Lexicons*, clearly presents the need for large-scale lexicons, and addresses the question of "How can we build the lexicons we need in the time available?" Almost every one of these papers discusses the difficulty of augmenting dictionary entries, whether machine-readable or not, with the information required by a natural language application, and the necessity of finding automatic methods for this augmentation. The section begins with a comprehensive overview of the field by Bran Boguraev that provides a useful context for the rest of the papers. It is followed by a paper from Bob Amsler describing the automated derivation of a lexicon from newspaper articles. The subsequent papers examine the advantages and drawbacks of different ways in which Machine Readable Dictionaries (MRDs) can be used for the automatic derivation of information for a computational lexicon. The first of these, a paper by Roy Byrd, focuses on adequately representing polysemous word senses, relying as much as possible on information that is automatically derived from machine readable dictionaries. Then Eva Hajíčová discusses a system that obtains grammatical information about verb argument structure from machine readable dictionaries. This is followed by a detailed description of an actual implementation, the impressive IIT lexicon project headed by Martha Evens. The Judy Klavans paper presents proposals for incorporating a semantic net derived from MRDs into information retrieval systems. Sue Atkins, Judy Kegl and Beth Levin carefully examine the elements of a verb's meaning that are reflected in its syntactic alternations, and how they are and are not captured by typical dictionary entries. Nicoletta Calzolari's paper, ten years after the Boguraev paper and after much work on deriving part of the information needed for a computational lexicon from MRDs, puts forward a number of issues which anyone aiming at building a very large multifunctional lexicon for NLP should pay attention to. The paper by Nancy Ide proposes a novel data base scheme especially designed for computational lexicons. The Sergei Nirenberg and Lori Levin paper introduces *constructions*, a class of linguistic phenomena that is especially difficult to deal with on a word by word basis. The paper by Petr Sgall discusses in detail the underlying theoretical approach for including syntactic information into lexical entries. This section concludes with an intriguing discussion of a new approach to representing semantic information by Helmut Schnelle.

This brings us to our third section, *The Acquisition and Use of Large Corpora* which is concerned with making the best possible use of the vast quantities of text that are now available electronically, clearly situated in Karen's statistical phase. The paper by Don Walker gives a broad overview of the field. It is followed by a paper by Doug Biber which deals with corpus design, a very important issue in the collection of a corpus. C.

M. Sperberg-McQueen discusses guidelines and standards in text representation. The paper by Bill Gale, Ken Church, and David Yarowsky describes significant advances in the application of statistical word probabilities to diverse problems in disambiguation. Both the Susan Armstrong-Warwick and Susan Hockey papers are aimed at the task of acquiring and disseminating text in electronic form. This section ends with a paper that discusses ways in which the corpora can be used, with Nick Belkin looking at principled ways of making on-line data available to scholars.

We end the book with a section that returns to the theme of system building, and discusses the integration of separate components, and issues involved in applying these techniques to other languages: *Topics, Methods, and Formalisms in Syntax, Semantics and Pragmatics*. Syntax has been an area of concern from the beginning, when the early work on machine translation revealed the necessity of valid syntactic information, "... long-distance dependencies, the lack of a transparent word order in languages like German, and also the need for a whole-sentence structure characterization to obtain properly ordered output, as well as a perceived value in generalization, led to the development of autonomous sentence grammars and parsers." In addition to highlighting the field's undying interest in parsing, this section also illustrates the more recent move towards cross-linguistic studies. We begin with a topical paper by Aravind Joshi that explores the applicability of statistical data derived from large corpora to the problem of syntactic parsing. He presents a hybrid approach which incorporates statistical data into a more conventional grammar based parsing system, signaling a recent trend towards the integration of statistical information with more traditional approaches. The next three papers are also primarily concerned with syntax, in a more and more global sense. While the Hobbs paper focuses on strategies for the syntactic parsing of English, the Nagao paper examines the applicability of English grammars and parsing algorithms to the processing of Japanese. The Rosner paper describes a state of the art formalism based on feature unification that can be applied cross-linguistically. The rest of this section concentrates on integrating syntactic information with other sources of information, such as semantics and pragmatics. The Joyce Friedman and Dave Warren paper describes their implementation of Montague grammar, which subsequently served as the springboard for much of the grammatico-logical phase. It is followed by the Martha Palmer, et al., paper, presenting a later implementation that used a less formal semantic representation as the basis for an integration of semantics and pragmatics. The following more recent paper, also a pragmatics paper, is by Cecile Paris and Vibhul Mittal, and describes a system implementation that requires a user model, in particular knowledge of the user's expertise, to appropriately tailor text generation. Finally, we end this section where we began, with a look at the use of statistical data derived from large corpora, this time for the comprehensive task of machine translation. Yorick Wilks discusses the IBM statistically based machine translation system which he portrays as having fallen somewhat short of its original goals. Similarly to Joshi, Wilks argues in favor of a hybrid approach that will incorporate statistical methods into more traditional systems. This is a positive note on which to end, showing that we can continue to build on the work of the past, incorporating first linguistics, then artificial intelligence and formal logic, and now even statistical methods, into our natural language processing systems, making them richer, more robust, and more reliable, and bringing them yet another step closer to that ultimate goal, of someday really understanding what a human is trying to communicate.

The process of putting the volume together has been very revealing of Don's personal touch, the inimitable stamp of courage and gallantry that was so uniquely his. Every author was kind and gracious to deal with, eager to make a contribution and to do everything possible to ease the task of the editors. And every author could relate some incident, some little story, some way in which Don had put them in his debt; a debt they delighted in repaying. This personal view of Don comes across very poignantly in *Don Walker: A Remembrance* by Jerry Hobbs and Barbara Grosz which follows this introduction. We can say unanimously, with the contributors to this volume, and his colleagues from Bellcore, "Our greatest debt is to Don Walker, whose incredible energy first brought us together and inspired us to make this practical [natural language] application a reality. His legacy will fire the imagination and influence the practice of computational natural language processing for ages to come."

In addition we would like to express our appreciation to all of the authors and to the many people who worked tirelessly to put this volume together in record time, and to an unusually flexible publisher, Giardini. We could never have made our deadline without the generous help of many University of Pennsylvania graduate students and staff who volunteered for proofreading and copyediting, such as Julie Bourne, Brett Douville, Dania Egedi, Nobo Komogata, Libby Levison, Michael Niv, Joseph Rosenzweig, Matthew Stone and Mike White. We would especially like to thank Tilman Becker, Alexis Dimitriadis, B. Srinivas, and David Yarowsky who worked into the small hours of the morning to resolve all of our latexing conundrums. Aravind Joshi very generously provided the necessary computing and printing resources. Allison Anderson did a sterling job as copy editor with an impossible overnight turnaround, and finally, our heartfelt gratitude goes to Carolyn Elken and Tarina Ayazi, who kept track of all the papers, and made it all happen. Carolyn, in particular, made latexing and copyediting changes to almost every paper, in a prodigious effort on her part that single-handedly enabled us to finish on time.

Martha Palmer  
May, 1994

# **Donald Walker: A Remembrance**

with contributions from

Robert Amsler, Woody Bledsoe, Barbara Grosz,  
Joyce Friedman, Eva Hajičová, Jerry R. Hobbs,  
Susan Hockey, Alistair Holden, Martha Palmer,  
Fernando Pereira, Jane Robinson, Petr Sgall,  
Karen Sparck Jones, Wolfgang Wahlster, and Arnold Zwicky

Don Walker had a vision of how natural language technology could help solve people's problems. He knew the challenges were great and would require the efforts of many people. He had a genius for bringing those people together.

For this introduction, a number of people who had known Don over the years were asked to write reminiscences. Though each person's story differed, a striking commonality emerged. It is remarkable how often Don was present at the key juncture in people's careers, and, in his understated, soft-spoken, low-key way, he did just the right thing for them. What Don did almost always involved bringing people together.

There is a story by Jorge Luis Borges in which someone travels all over India attempting to discover the nature of a very wise man, through the subtle but profound influence he had had on the people he met. Reading the reminiscences of Don and seeing the impact he had on people's lives reminded us of that story.

Don organized research teams. He was often instrumental in matching people with positions in laboratories located far from his own as well as in groups he managed. Several people attribute to Don significant help in starting their careers or in finding the funding that supported their most productive period of research. He often gave essential advice or provided the key opportunity that led to a fruitful new direction in someone's career. Don knew who was doing everything, how to get in touch with them, and how to facilitate the appropriate actions that matched the person with the opportunity.

This was first evidenced in the group he built at Mitre, after a sojourn at Rice University and a year at MIT. A psychologist by training, Don was one of the first to recognize the relevance of contemporary MIT linguistics to computational processing of language and in 1963 at MITRE he found Air Force support for that application. The work at MITRE was on the problem of formalizing grammars for computer application to obtain "deep structures" of grammatical sentences by means of inverse transformations—transformational grammar. It resulted in a large computer program, described in internal MITRE documents, and presented to the 1965 AFIPS conference. The group he assembled was strikingly interdisciplinary, especially for its time. At one stage or another, it included a large number of linguists, philosophers, psychologists, mathematicians, and computer scientists. Linguistics students from MIT played a key

role, often spending their summers at MITRE. Among them were many young scholars such who have since become leaders in their fields, and remain grateful to Don for his help when they were starting out. Some of the people who worked at Mitre during this period were Ted Cohen, Tom Bever, Paul Chapin, Ted Cohen, Bruce Fraser, Joyce Friedman, Mike Geis, Dick Glantz, Ted Haines, Barbara Hall (Partee), Steve Isard, Dick Jeffrey, Jackie Mintz, Stan Peters, Peter Rosenbaum, Haj Ross, Tim Scanlon, Sandy Shane, Howard Smokler, and Arnold Zwicky.

Don was an amazing project leader at MITRE. He encouraged the widest range of explorations, gently but firmly critiqued everything everyone did, and somehow coped with the corporation and military brass. Yet he was always self-effacing and almost apologetic about his accomplishments.

Don went on from Mitre to SRI, where in the 1970s he built the natural language group there that continues today as one of the premier natural language research groups in the world. Barbara Grosz recalls how he gave her her first AI job, even though she had yet to pick a thesis topic, let alone finish a Ph.D. In doing so, he took a risk of a magnitude that she fully appreciated only years later when she was hiring research associates. It was not a unique gamble; Don was often credited at Mitre with identifying good people before their reputations were widely established and he continued this process at SRI.

Jerry Hobbs recalls meeting Don at the ACL conference in Boston in 1975. Don greeted him so enthusiastically that he applied for a job at SRI. Little did he know that Don greeted everyone that way.

Don was also a wonderful group leader at SRI, not least because he so thoroughly integrated the personal and professional. Don appreciated and cared about the whole person of each member of the group, and nurtured them. They were as much family as research group, working hard and arguing hard, but appreciating one another and truly a team. They bonded closer in crises in a way few groups achieve. Don knew his people and led them to work together through the best and the worst. He demonstrated that one need not compromise personal warmth or care for individuals to have a top-rate research group. He also continued to provide guidance for members of his group even after they were in different labs. He often worked behind the scenes in many places and at many crucial junctures to make sure that their work was recognized. Don would never admit to these efforts, but he smiled in a certain way when asked. Asking was the closest one could come to saying thanks, and the smile the only sign that he'd welcomed the appreciation.

Don not only brought people together; he continued to bring disciplines together. In forming the SRI natural language group, he again built up an interdisciplinary group of researchers long before "cognitive science" was a phrase, and long before the need for interdisciplinary approaches to the study of language was taken for granted. He firmly believed that interdisciplinary work was the only way to go; to succeed at building the kind of computer systems they were after, AI researchers would have to be informed by linguistics, philosophy, psychology, and sociology. He constructed an environment that freed the people in his group to concentrate on their research, and he communicated his own excitement and enthusiasm for what could be done in this new interdisciplinary, unpredictable field where human language and computers meet.

Don played a central role in the organization and operation of the Association for Computational Linguistics (ACL), the International Conference on Computational

Linguistics (Coling), the International Joint Conference on Artificial Intelligence (IJCAI), and the American Association for Artificial Intelligence (AAAI). He was Secretary/Treasurer for most of them, some for many decades. Expressing a theme that appeared again and again in the reminiscences, Wolfgang Wahlster, Conference Chair of IJCAI-93, said, "Don guided me like a father through the complex IJCAI world." This feeling is one every ACL and IJCAI official through the years has shared.

When Don took over from Hood Roberts as Secretary-Treasurer of ACL, the organization was relatively small. It grew, and Don's job grew with it, in size and complexity. Over the years, ACL officials have had a glimmer of how many concerns Don had to deal with and how much he did, and they have been vaguely aware that there were all sorts of other things he was managing behind the scenes. But only now have the scale and variety of these responsibilities become apparent. In a report about her recent visit to Don and Betty's house in Bernardsville, Karen Sparck Jones said, "Betty told me that in looking for houses when they moved to New Jersey one of the important considerations was that there had to be a good big basement for the ACL office. Sitting in it and talking with Don and Betty about the gritty details of ACL's finances was a direct encounter with what running something like ACL has implied for Don and thus how much he has contributed to building it up to the first-class society it is." Evident in all this is not only that Don did an enormous amount of work keeping the ACL show on the road and improving its act all the time, but also that he had continual concern and respect for all the different parties in the ACL and for the ACL's wider interests in the community and, especially, the international community.

Don was Program Chair of the first IJCAI and General Chair of the next IJCAI. Alistair Holden, who was General Chair of the first IJCAI, reports that as the General Chair of the second conference Don "organized IJCAI as it is today; an independent body governed by trustees who come from the international AI community. . . [He] provided the continuity and organization that have led to the present IJCAI success." As Secretary-Treasurer of IJCAI for many years, Don provided invaluable guidance to successive Conference Chairs and Boards of Trustees. Recently, the Principles of Knowledge Representation and Reasoning conference, incorporated March 1993, adopted a governing structure that closely mirrored IJCAI's. It is no small tribute to Don that a structure he formulated more than twenty years ago is being copied today.

In recognition of Don's many contributions to the IJCAI organization and his very great service to the international AI community, the IJCAI, Inc. Board of Trustees renamed the IJCAI Distinguished Service Award at IJCAI-93. The award is now the Donald E. Walker Distinguished Service Award. Daniel G. Bobrow, the first recipient of the newly named award and a former President of AAAI, remarked on accepting the award that Don had provided a model of service which everyone in the field could look up to.

Don's well-known organizational skills led Raj Reddy to invite him to participate in formative discussions of the AAAI (on the way to IJCAI-79) and designated him the first Secretary-Treasurer. Don's knowledge of financial procedure was invaluable in those early days, as was his considerable memory for the details of the charter, California law, and resolutions passed in Council meetings. Don was so absolutely dependable and trustworthy that it was not until well after his term was up that anyone thought that the treasury of the AAAI needed any protection from an incumbent Treasurer.

Don was always fair; he carried his sense of what was right and fair to the complex discussions of arrangements between AAAI and IJCAI; his considerable efforts and negotiation skills were important to establishing a good working relationship between the two organizations.

A listing of Don's service roles does not make apparent the depth of his contribution, for that goes far beyond the day-to-day management he often provided. Don's service mattered so much because it was always in the interests of a larger goal that he cared passionately about. When he participated in the organization of a new conference or the founding of an organization, it was not for its own sake (or because he didn't already have enough irons in the fire), but because he thought it was necessary to further an important research goal. The goals, as well as the conferences and organizations, were always integrative. IJCAI brings together researchers from around the world and across the fields of AI; Don constantly reminded the Trustees and Advisory Board of the importance of safeguarding both dimensions of its diversity. Coling is likewise international, and, with much guidance from Don, ACL has also evolved into an international body. Having helped to foster the now-flourishing European Chapter of the ACL, he looked toward a Chapter on the Pacific Rim. Don was always inclusive, never exclusive.

Don brought people together in other ways. He organized a fund and talked to the right people to give Eastern Europeans the right to visit their colleagues abroad. Petr Sgall remembers his first long personal conversation with Don at the romantic, medieval Hungarian castle of Visegrad. "Our conversation touched perhaps everything from text retrieval to topical issues of world politics, but Don's main concern was to find maximally effective ways of overcoming the disadvantages of researchers in Communist countries. I then saw that Don's effort towards this aim were of much larger dimensions than what I could imagine. Don's conception of international cooperation went far beyond generous support for individual stays. For example, thanks to his initiative, such valuable sources of information as the journal *Computational Linguistics* reaches dozens of colleagues who, due to the unfavorable conditions in their countries, would not otherwise be able to get it."

The computational linguistics community in Eastern Europe benefitted greatly from Don's efforts toward their integration into the international research community, especially the initiation of the international fund which gave them all the advantages of ACL membership. Eva Hajičová recalls, "When Don called me at the beginning of 1982 from California . . . and asked me whether such an action would be good for us, I just could not believe my ears. It lasted a while before I realized the depth of his offer and became able to express our gratitude and to think of the technicalities."

In tribute to Don's efforts to include as many people as possible in the computational linguistics endeavor he found so exciting, and his and Betty's longstanding contributions to the ACL, the ACL Executive Committee established the Don and Betty Walker Student Fund. This fund enables students to attend ACL meetings. His legacy will live on, bringing together people who never had the good fortune to know him.

In recent years, Don's research and his organizational efforts took a new turn. Corpus-based natural-language processing, which at present may be the fastest-growing and most exciting area of computational linguistics, owes a particular debt to Don's vision and leadership. He recognized the importance of working on *real* natural-language corpora and other linguistic knowledge resources before that was in vogue, and he

developed and sustained successful efforts to organize the collection and annotation of databases for corpus-based research. In particular, he played crucial leadership roles in the ACL Data Collection Initiative, which has now made available a variety of corpora to the international research community, and in the Text Encoding Initiative, which has brought together international organizations concerned with textual research to create a standard for encoding machine-readable text. Through the Text Encoding Initiative, Don became interested in humanities computing and its approaches to working with large textual databases. He soon realized how much that community and computational linguistics could benefit from interaction with each other and in his characteristic way went on to foster those interactions. His influence and support can also be seen in the European Corpus Initiative, the Linguistic Data Consortium, and the Consortium for Lexical Research, all of which are actively contributing to the variety and quantity of corpora and lexical resources available to the research community in natural-language and speech processing, linguistics and the humanities. All of this was in the service of a long-term vision he referred to as “The Ecology of Language”, that is, the attempt to characterize the contexts in which people use language, and thus in which natural language technology can be made useful to people. This theme is reflective of his aims during his whole career.

We can't resist one more anecdote that shows that he was able not only to bring people together, but, at least in one instance, to keep them together for a few more hours. Paul Martin held a large party to celebrate receiving his Ph.D., and after midnight it was getting noisy enough that the neighbors called the police. When they came, Don told Paul not to worry. He'd handle it. He went out to talk to them, in his inimitable, reasonable, soft-spoken manner. No one knows what he said, but the police left and the party continued.

The delight that Don took in his work and in the people he worked with was infectious. Many times he would try to share that delight with friends who were strangers to the field. His face would light up; his hands would orchestrate his attempts to make them understand the wonder of it all. Often he would search in vain for a word that could convey how he felt about the field, the people, and the ideas, and he would end up saying something like “It's just so . . . so . . .” and fall back with a sigh to “elegant”, or “amazing”, extending his hands as if to shape the elusive message.

The ancient Chinese reserved a special place in heaven for people who built bridges. Don Walker built bridges.

## **SECTION I**

### **THE TASK OF NATURAL LANGUAGE PROCESSING**

# Natural Language Processing: A Historical Review\*

Karen Sparck Jones  
Computer Laboratory, University of Cambridge  
*e-mail: kn11@phx.cam.cl*

## **Abstract**

This paper reviews natural language processing (NLP) from the late 1940's to the present, seeking to identify its successive trends as these reflect concerns with different problems or the pursuit of different approaches to solving these problems and building systems as wholes. The review distinguishes four phases in the history of NLP, characterised respectively by an emphasis on machine translation, by the influence of artificial intelligence, by the adoption of a logico-grammatical style, and by an attack on massive language data. The account considers the significant and salient work in each phase, and concludes with an assessment of where we stand after more than forty years of effort in the field.

## **1 Introduction**

At the ACL Conference in 1987 Don Walker, Jane Robinson and I were talking about when we began in NLP research. Fred Thompson told us he began in 1954 and others, like Martin Kay, started out too in the fifties. Work in the field has concentrated first on one problem, then on another, sometimes because solving problem X depends on solving problem Y but sometimes just because problem Y seems more tractable than problem X. It is nice to believe that research in NLP, like scientific research in general, advances in a consolidating way, and though there may be more faith than substance in this, we can certainly do NLP now we could not do in the fifties. We may indeed be seduced by the march of computing technology into thinking we have made intellectual advances in understanding how to do NLP, though better technology has also simply eliminated some difficulties we sweated over in earlier years. But more importantly, better technology means that when we return to long-standing problems they are not always so daunting as before.

Those, like Don, who had been around for a long time, can see old ideas reappearing in new guises, like lexicalist approaches to NLP, and MT in particular. But the new costumes are better made, of better materials, as well as more becoming: so research is not so much going round in circles as ascending a spiral, if only a rather flat one. In reviewing the history of NLP, I see four phases, each with their distinctive concerns and styles. Don, in one way or another and like all of us, to some extent moved in

---

\*The material in the earlier part of this paper is taken from my article "Natural language processing: an overview", in W. Bright (ed.) *International encyclopedia of linguistics*, New York: Oxford University Press, 1992, Vol. 3, 53-59.

time to the current beat. But it is noteworthy that in the push he made for linguistic resources, like corpus collections, he not only significantly promoted what I have called the data-bashing decade that is now with us, but also returned to what was a major concern in the first period of NLP research: building the powerful and comprehensive dictionaries that serious NLP applications, like MT, need.

I define the first phase of work in NLP as lasting from the late 1940s to the late 1960s, the second from the late 60s to the late 70s and the third to the late 80s, while we are in a clear fourth phase now.

## 2 Phase 1: Late 1940s to Late 1960s

The work of the first phase was focused on machine translation (MT). Following a few early birds, including Booth and Richens' investigations and Weaver's influential memorandum on translation of 1949 (Locke and Booth, 1955), research on NLP began in earnest in the 1950s. Automatic translation from Russian to English, in a very rudimentary form and limited experiment, was exhibited in the IBM-Georgetown Demonstration of 1954. The journal *MT (Mechanical Translation)*, the ancestor of *Computational Linguistics*, also began publication in 1954. The first international conference on MT was held in 1952, the second in 1956 (the year of the first artificial intelligence conference); at the important Washington International Conference on Scientific Information of 1958 language processing was linked with information retrieval, for example in the use of a thesaurus; Minsky drew attention to artificial intelligence; and Luhn provided auto-abstracts (actually extracts) for one session's papers. The Teddington International Conference on Machine Translation of Languages and Applied Language Analysis in 1961 was perhaps the high point of this first phase: it reported work done in many countries on many aspects of NLP including morphology, syntax and semantics, in interpretation and generation, and ranging from formal theory to hardware.

This first phase was a period of enthusiasm and optimism. It is notable not only because those engaged attacked a very difficult NLP task, and so encountered the problems of syntactic and semantic processing, and of linguistic variety, in all their force; they were seeking to use a new tool, computers, for non-numerical, data-processing purposes when data-processing itself was not well established. It is essential to remember how primitive the available computing resources were. This was the era of punched cards and batch processing. There were no suitable higher-level languages and programming was virtually all in assembler. Access to machines was often restricted; they had very limited storage, and were extremely slow. Plath (1967) reports processing speeds like 7 minutes for analysing long sentences, even with the most advanced algorithms and on the best machines then available. Vast amounts of programming effort were devoted to bit-packing to save space and time. It is remarkable how much was done with such poor resources, for example in grammar and lexicon building: some of the grammars and dictionaries of the early 1960s were very large even by current standards.

Research in this period was thoroughly international, with considerable activity in the USSR as well as in the USA and Europe, and some in Japan. US grant funding

increased after Sputnik 1, but the work had begun before. Russian and English were the dominant languages, but others, including Chinese, were involved (Booth, 1967; Hutchins, 1986).

Though the period ended under the cloud of the 1966 ALPAC Report, (ALPAC, 1966; Hutchins, 1986), most of those engaged were neither crooks nor bozos. Many came to NLP research with a background and established status in linguistic and language study, and were motivated by the belief that something practically useful could be achieved, even though the strategies adopted were crude and the results not of high quality. The first major question was whether even to obtain only limited results, principled methods based on generalisation were required, or whether ad hoc particularisation would suffice. The second issue was the relative emphasis to be placed, in either case, on syntax and on semantics. The third problem was the actual value of the results, especially when balanced against pre- or post-editing requirements.

The main line of work during this period can be summarised as starting with translation as lookup, in dictionary-based word-for-word processing. The need to resolve syntactic and semantic ambiguity, and the former in particular because it is not open to fudging through the use of broad output equivalents, led to ambiguity resolution strategies based on local context, so dictionary entries became in effect individual procedures. Semantic resolution involved both specific word, and semantic category, collocation. But long-distance dependencies, the lack of a transparent word order in languages like German, and also the need for a whole-sentence structure characterisation to obtain properly ordered output, as well as a perceived value in generalisation, led to the development of autonomous sentence grammars and parsers.

Most of the NLP research done in this period was focused on syntax, partly because syntactic processing was manifestly necessary, and partly through implicit or explicit endorsement of the idea of syntax-driven processing. The really new experience in this work, and its contribution to linguistics in general, came from recognising the implications of computing represented by the need not only for an explicit, precise, and complete characterisation of language, but for a well-founded or formal characterisation and, even more importantly, the need for algorithms to apply this description. Plath's account (1967) of NLP research at Harvard shows this development of computational grammar with its lexicon and parsing strategy very clearly. But as Plath also makes clear, those concentrating on syntax did not suppose that this was all there was to it: the semantic problems and needs of NLP were only too obvious to those aiming, as many MT workers were, at the translation of unrestricted real texts like scientific papers. The strategy was rather to tackle syntax first, if only because semantic ambiguity resolution might be finessed by using words with broad meanings as output because these could be given the necessary more specific interpretations in context.

There were however some workers who concentrated on semantics because they saw it as the really challenging problem, or assumed semantically-driven processing. Thus Masterman's and Ceccato's groups, for example, exploited semantic pattern matching using semantic categories and semantic case frames, and indeed in Ceccato's work (1967) the use of world knowledge to extend linguistic semantics, along with semantic networks as a device for knowledge representation.

MT research was almost killed by the 1966 ALPAC Report, which concluded that MT was nowhere near achievement and led to funding cuts especially in the most active

country, the USA, even though it recommended support for computational linguistics. But it is important to recognise what these first NLP workers did achieve. They recognised, and attempted to meet, the requirements of computational language processing, particularly in relation to syntactic analysis, and indeed successfully parsed and characterised sentences. They investigated many aspects of language, like polysemy, and of processing, including generation. They addressed the issues of overall system architectures and processing strategies, for example in direct, interlingual or transfer translation. They began to develop formalisms and tools, and some influential ideas first appeared, like the use of logic for representation (cf. Yngve, 1967). Some groups were also established, developing resources like grammars and gaining experience, as at the Rand Corporation. There was indeed enough knowhow by now for some textbooks, like Hays (1967).

There was little work, on the other hand, on some important problems that have since attracted attention, like anaphor resolution, since though text was being translated it was treated as a sequence of independent sentences, or on the function of language, since the work was mainly on single-source discourse. There was little attempt to incorporate world knowledge, and to relate this non-linguistic knowledge to linguistic knowledge, though some world knowledge was smuggled in under the heading of semantics. The belief, or challenge, was that one could get far enough with essentially linguistic, and therefore shallow, processing not involving reasoning on world models. The research of this period did not produce any systems of scope or quality, though by the end of the 1960s there were MT production systems providing output of use to their customers (Hutchins, 1986). There was more merit in the work of the period, and more continuity, through individuals, with later effort, than subsequent myths allow, though the early literature was inaccessible and little used. But perhaps the best comment is Bledsoe's at the International Joint Conference on Artificial Intelligence of 1985 (Bledsoe, 1986) on the value, for artificial intelligence as a whole, of the early MT workers' head-on attempt to do something really hard.

Work on the use of computers for literary and linguistic study also began in this period, but it has never been closely linked with that in NLP, though some common concerns have become more prominent recently.

### 3 Phase 2: Late 1960s to Late 1970s

The second phase of NLP work was artificial intelligence (AI) flavoured, with much more emphasis on world knowledge and on its role in the construction and manipulation of meaning representations. Pioneering work influenced by AI on the problems of addressing and constructing data or knowledge bases began as early as 1961, with the BASEBALL question-answering system (Green et al, 1961). The actual input to these systems was restricted and the language processing involved very simple compared with contemporary MT analysis, but the systems described in Minsky (1968), and Raphael's SIR in particular, recognised and provided for the need for inference on the knowledge base in interpreting and responding to language input.

Woods et al.'s LUNAR (Woods, 1978) and Winograd's SHRDLU (Winograd, 1973) were the natural successors of these systems, but they were widely seen at the time as representing a step up in sophistication, in terms of both their linguistic and their task-

processing capabilities. Though differing in many ways they shared a procedural style and were perceived as having an overall coherence as systems and a genuinely computational character. The dominant linguistic theory of the late 1960s, transformational grammar, was seen both as fundamentally unsuited to computation and particularly analysis, even though TG was formally oriented and there was at least one serious transformational parser, and as offering nothing on semantics, which had to be tackled for any actual NLP system. The computational confidence illustrated by Woods' and Winograd's work, and the range of experiment it promoted, while drawing on previous work, is well shown by the varied research reported in Rustin (1973).

The view that current linguistics had nothing to contribute, and the feeling that AI was liberating, were also apparent in Schank's work (1980), which explicitly emphasised semantics in the form of general-purpose semantics with case structures for representation and semantically-driven processing. The community's concern, illustrated by Winograd and Schank alike, with meaning representation and the use of world knowledge then became an argument, reflecting a widespread feeling in AI stimulated by Minsky's promulgation of frames (Minsky, 1975), for the use of a larger scale organisation of knowledge than that represented in NLP by verb case frames or propositional units: this large-scale organisation would characterise the different relationships between the elements of a whole universe of discourse, and would support the inferences, including default inferences, needed especially in interpreting longer discourse and dialogue. NLP would deliver deep representations integrating and filling out individual inputs to form a whole constituting an instantiation of a generic world model. Schank's arguments for the Yale group's use of more event-oriented scripts developed this line in the context of earlier work by linking individual propositional case frames with the larger structures via their semantic primitives (cf. Cullingford, 1981). Semantic networks (Bobrow and Collins, 1975; Findler, 1979) were similarly proposed as a third variant on this theme, offering a range of options from associative lexical networks only weakly and implicitly embodying world knowledge to alternative notations for frames. These types of knowledge representation linked NLP with mainstream AI, and their descriptive and functional status, for example in relation to logic, was and has remained a matter for debate.

Semantic primitives seen, as in Schank's Conceptual Dependency Nets (Schank, 1975), as having a representational and not just a selective role also appeared to fit naturally with the need to capture underlying conceptual relations and identities in discourse processing, particularly for types of material or tasks where fine distinctions do not figure. Their status too was a matter for controversy, but they have continued in use, supplemented by or sometimes in the form of domain-specific categories, in application systems. They have also had a significant role, in the more conventional form of selectional restrictions, even when semantic driving has been abandoned.

The general confidence of those working in the field, and the widespread belief that progress could be and was being made, was apparent on the one hand in the ARPA Speech Understanding Research (SUR) project (Lea, 1980) and on the other in some major system development projects building database front ends. Several of the SUR projects were ambitious attempts to build genuinely integrated systems combining top-down with bottom-up processing, though unfortunately the best performing system against the target measurements was the least theoretically interesting.

The front end projects (see, e.g., Hendrix et al., 1978) were intended to go signifi-

cantly beyond LUNAR in interfacing to large autonomous (and therefore not controlled) databases, and in being more robust under the pressures of 'ill-formed' input; and the confidence on which they were based drove other work including that on the first significant commercial front end, INTELLECT (Harris, 1984). But these projects unfortunately also showed that even an apparently straightforward, and perhaps the simplest because naturally constrained, NLP task was far more difficult than it seemed to be. NLP workers have been struggling ever since on the one hand with the problems of constructing general-purpose transportable front ends and of providing for the acquisition of application-specific knowledge, and on the other of handling the user's real needs in dialogue. The former led to the development of modular architectures, general-purpose formalisms, and toolkits, typically for supplying a specialised lexicon, semantics, and domain and database model on top of standard syntax, following the sublanguage approach which had been pioneered for text processing by Sager's NYU group (in Kittredge and Lehrberger, 1982), but sometimes supplying a specialised syntax as well. The latter stimulated research on the identification of the user's beliefs, goals and plans which is also and more fully needed for dynamic and extended interaction with expert systems for consultation and command, where the system's responses should be cooperative.

The need to identify the language user's goals and plans was early recognised by the Yale group, and has become a major trend in NLP research since, along with a more careful treatment of speech acts. Work on interactive dialogue in particular, from the second half of the 70s, has emphasised the communicative function of language, and the indirect function and underlying meaning, as well as direct function and surface meaning, of linguistic expressions. At the same time work on discourse understanding in the 70s, whether on single-source texts like stories or reports, or on dialogue, stimulated research on anaphor resolution and on the construction, maintenance and use of discourse models not relying only on prior scenarios like scripts; and some useful progress was made with the development of notions of discourse or focus spaces and of resolution algorithms tied to these (Joshi et al., 1981; Brady and Berwick, 1983; Grosz et al., 1986).

#### 4 Phase 3: Late 1970s to Late 1980s

It was nevertheless apparent by the early 1980s that it was much harder to build well-founded, i.e., predictable and extensible, NLP systems even for heavily restricted applications than had been supposed, and that systems for more challenging applications in terms of processing tasks or discourse domains could not generally be built in an ad hoc and aggregative way, though claims were made for this as a possible strategy for utilitarian MT, given enough investment of effort.

If the second phase of NLP work was AI-flavoured and semantics-oriented, in a broad sense of "semantic", the third phase can be described, in reference to its dominant style, as a grammatico-logical phase. This trend, as a response to the failures of practical system building, was stimulated by the development of grammatical theory among linguists during the 70s, and by the move towards the use of logic for knowledge representation and reasoning in AI. Following augmented transition networks as computational grammars in a theoretical as well as practical sense, linguists developed a whole range of grammar types, for example functional, categorial and generalised phrase structure, which, because they are oriented towards computability as an abstract

principle, are also relevant to actual parsing, particularly since they also tend to have a context-free base supporting efficient parsing algorithms. The emphasis was also on a declarative approach and on unification as the fundamental process, which fitted naturally with a general trend in computing in this period associated with, for example, the growth of logic programming. The processing paradigm, for analysis in particular, was therefore syntax-driven compositional interpretation into logical forms.

Computational grammar theory became a very active area of research linked with work on logics for meaning and knowledge representation that can deal with the language user's beliefs and intentions, and can capture discourse features and functions like emphasis and theme, as well as indicate semantic case roles. The issues in this approach are those both of reflecting the refinements of linguistic expressions in indicating time and mood or conveying presuppositions, and of preserving cohesive and coherent discourse structure. However the belief that the grammatico-logical route is the right, because principled, way to go did lead by the end of the 80s to the development of powerful, general-purpose processors, of which SRI's Core Language Engine (Alshawi, 1992) can be taken as an exemplar. These processors could be used to support application systems with at least as much operational power as ones based on less absolutist views, for example on transition nets and frames, and with more potential for superior performance when challenged.

The grammatico-logical approach was also influential in some other ways. It led to the widespread use of predicate calculus-style meaning representations, even where the processes delivering these were more informal than the purist would wish. It also led, when taken with the challenge of building effective systems for, e.g., database query, to a shift in the meaning of "semantic" and "pragmatic" and to changes in the distribution of effort over the system as a whole. Semantic interpretation, given basic lexical data, concentrated on e.g., quantifier interpretation, and the full meaning of expressions was taken to be supplied by reference to the pragmatic context, subsuming both the prior discourse context and the application's domain or world model.

All together, the period can be seen as one of growing confidence and consolidation, partly encouraged by the general enthusiasm associated with the Fifth Generation enterprise, but also well-justified by the ability to build better systems, itself reflected in the beginning of the ACL's series of Applied NLP Conferences.

In relation to the central concerns of NLP, consolidation is most evident in syntax, the area in which, from an historical point of view, most progress has been made. By the end of the 1980s, practical system builders could take advantage of relatively well-understood forms of grammar and parsing algorithm, and also sometimes of large actual grammars and bodies of software, like those of the Alvey Natural Language Tools (cf. Briscoe et al., 1987). At the same time, other operational systems joined SYSTRAN and METEO (cf. Hutchins and Somers, 1992) in NLP applications, which now addressed a range of tasks including, e.g., message processing as well as translation, and commercial systems were both offered and purchased, especially for database query (cf. Engelien and McBryde, 1991). Research and development extended world-wide, notably in Europe and Japan, aimed not only at interface subsystems but at autonomous NLP systems, as for message processing or translation. However there was to some extent a division in this period between those focusing on principles and those focusing on practical applications, who did not always follow the formalist, grammatic-logician

line but exploited whatever conceptual apparatus was to hand, like case and domain frames.

The revival of MT was a significant feature of this period, in which European and Japanese interest played a major part. The European Commission both used production systems based on customised pragmatism and promoted the Eurotra research project on multi-lingual translation within a common, well-defined transfer framework. There were several active Japanese teams, with some translation products in the market (Nagao, 1989). Much of the MT work done assumed that something at least useful and perhaps more could be provided, particularly for specific applications, with or without editor or user participation in the translation process; and it reflected the current state of NLP in grammar choices and the use of modular system architectures.

On the research side, the period was notable for a growth of interest in discourse, and it saw the first serious work on generation, especially multi-sentence text generation. There were two sides to the interest in discourse, which came together in the context of interactive, dialogue systems, for instance for advice giving, where the need for cooperative system responses implies modelling of the participants' beliefs, goals and plans, and can naturally lead to the production of paragraph-length output, for instance in providing explanations. Work on user modelling, as illustrated in Kobsa and Wahlster (1989), was one strand in research on language use intended for active communicative purposes and on discourse structure as related to such purposes (Cohen et al., 1990). At the same time, as e.g., McKeown (1985) showed, rhetorical schemas could be used as convenient recipes for producing communicatively effective, as well as linguistically coherent, text.

From the point of view of NLP as a whole on the other hand, there was more novelty in the connectionist approaches explored in this period, implying a very different system architecture from the conventional modular one (cf. Rumelhart et al., 1986). This work, though not directly absorbed into the mainstream, can be seen as one source, via the idea of probabilistic networks, for the present interest in statistically-flavoured NLP.

The final trend of the 80s was a marked growth of work on the lexicon. This was stimulated by the important role the lexicon plays in the grammatico-logical approach and by the needs of multi-lingual MT, and also by the problems of transportability, customising and knowledge acquisition in relation to individual applications. The first serious attempts were now made to exploit commercial dictionaries in machine-readable form, and this in turn led to the exploitation of text corpora to validate, enhance or customise initial lexical data, research made much easier by the rapidly increasing supply of text material. This last trend can be seen now to be giving the current fourth period of NLP its dominant colour.

## 5 Phase 4: Late 1980s Onward

Thus the last few years have seen a conspicuous move into statistical language data processing, so much so that this phase can perhaps be labelled the massive data-bashing period. Work on the lexicon has in part concentrated on the development of suitable general formalisms for expressing lexical information, closely tied to the way this is applied through operations on feature systems in syntactic and semantic processing, and taking advantage of AI experience in knowledge representation by viewing the lexicon

as a terminological knowledge base. But this work has been supported by notable initiatives in data gathering and encoding, and has encouraged a surge of interest in the use of corpora to identify linguistic occurrence and cooccurrence patterns that can be applied in syntactic and semantic preference computation. Probabilistic approaches are indeed spreading throughout NLP, in part stimulated by their demonstrated utility in speech processing and hence sometimes advocated not just as supports, but as substitutes, for model-based processing.

The rapid growth in the supply of machine-readable text has not only supplied NLP researchers with a source of data and a testbed for e.g., parsers. The flood of material has increased consumers' pressure for the means of finding their way round in it, and has led both to a new focus of NLP research and development in message processing, and to a surge of effort in the wider area of text processing which deals with the identification of the key concepts in a full text, for instance for use in text retrieval (cf. Jacobs, 1992). Thus NLP, earlier not found to be sufficiently useful for document retrieval based on abstracts, may contribute effectively to searching full text files. All of this work has encouraged the use of probabilistic tagging, originally applied only in data gathering, and the development of shallow or robust analysers. In this context, NLP workers have also been forced to handle more than well-formed individual sentences or well-mannered ellipses and to deal, for instance, with the variety of proper names.

The interest in text, as well as in improving the scope and quality of interfaces, has also promoted work on discourse structure, currently notable for the interaction between those approaching the determination and use of discourse structure from the point of view of computational needs and constraints, and those working within the context of linguistics or psycholinguistics.

A further major present trend can be seen as a natural outcome of the interaction between consumer (and funder) pressures and the real as well as claimed advances in NLP competence and performance made during the 1980s. This is the growth of serious evaluation activities, driven primarily by the (D)ARPA conferences (cf. HLT, 1993) but also reflecting a wider recognition that rigorous evaluation is both required and feasible when systems are solid enough to be used for non-trivial tasks (Galliers and Sparck Jones, 1993). Designing and applying evaluation methodologies has been a salutary experience, but the field has gained enormously from this, as much from learning about evaluation in itself as from the actual, and rising, levels of performance displayed. However evaluation has to some extent become a new orthodoxy, and it is important it should not turn into an ultimately damaging tuning to demonstrate prowess in some particular case, as opposed to improving the scientific quality of work in the field and promoting community synergy.

These evaluation initiatives have nevertheless focused attention on the challenge of NLP tasks involving operations on a large scale, like text retrieval from terabytes of material, and the nature of the specific tasks chosen has also had a stimulating effect in cutting across established boundaries, for instance by linking NLP and information retrieval. More importantly, the (D)ARPA conferences have helped to bring speech and language processing together, with new benefits for NLP from the improvements in speech processing technology since the SUR programme of the 1970s. These improvements are indeed more generally promoting a new wave of spoken language system applications, including ones involving translation, already demonstrated for limited do-

main inquiry systems and proposed in a much more ambitious form in the Verbmobil project (Kay et al., 1991; Siemens, 1991).

Finally, this period has seen a significant, new interest in multi-modal, or multi-media, systems. This is in part a natural response to the opportunities offered by modern computing technology, and in part an attempt to satisfy human needs and skills in information management. But whether combining language with other modes or media, like graphics, actually simplifies or complicates language processing is an open question.

## 6 Where We Are Now

Reviewing developments in the field as a whole over the last forty years, and what has been achieved, we find first, that the implications of computation in terms of the need for explicit data detail, proper process specification, and appropriate and adequate formalisms are now understood even if sometimes, as in the discourse area, it is too often taken for granted that outline theories can be translated into viable programs. The enormous improvements in machine technology have also meant, very usefully, that it is less essential than it was to worry about proliferations of alternatives during processing, while at the same time, whatever the attractions of cognitively convincing approaches, NLP can be well done in a purely engineering spirit. Moreover while major systems rest on person-decades of experience and effort, it is now possible, with present computing resources, to 'run-up' surprisingly powerful systems and to conduct impressively large experiments in a matter of months or even weeks.

In terms of what language processing requires, and specifically general-purpose language processing, most progress has been made in the area of syntax, where we have effective means of grammar characterisation and useful techniques like chart parsing. More generally, workers in the field now have a stock of conceptual tools, like case and domain frames, and enough experience of using them to put together a system or interface subsystem for many experimental or developmental purposes and even, for suitably restricted tasks or limited output expectations, for regular operational production. Performance can nowadays, moreover, be improved by exploiting probabilistic information. Advances in low-level speech processing have meant not only that performance in speech recognition without language understanding (as for dictation) is advancing, but that it is now possible to look for speech understanding systems with language processing capabilities not too far behind those for systems with typed input.

It is nevertheless the case that the most effective current systems, from the point of view of language understanding, are either those with the most limited domains or those with the least demanding tasks. The former include both systems based on putatively general-purpose machinery, customised in a tidy way, and systems essentially designed for given applications. In either case, though the tasks undertaken are not trivial, the systems operate within narrow bounds, for instance in relation to providing explanatory responses in dialogue, and are in general extremely brittle. Moreover while customising may be easier from a solid all-purpose base, there is so far little evidence for large performance gains for this rather than from the ad hoc approach. Overall, the challenge of taking the necessary step from a focused experiment or even convincing prototype to a full-scale rounded-out NLP system has not been overcome. Nagao's (1989) illustrations

of comparable translation performance for different systems is a salutary reminder of how far NLP has to go.

Nagao's examples, however, by showing how different translations may be equally acceptable, also emphasise the need for evaluation in user contexts, which is the key problem for the less demanding tasks, like document retrieval, where shallow processing may suffice but it is hard to show whether natural performance limits have been reached. Again, while highly modular architectures have been widely accepted, there are still major problems for all but the very limited or most tolerant applications, in determining the distribution of information and effort between the linguistic and non-linguistic elements in a system, and between the general-purpose and domain-specific components. Moreover, while NLP workers have enlarged their immediate fields and have begun, in particular, to escape from individual sentences and to handle larger wholes in dialogue and extended text, there are important language-using functions, or tasks, like summarising, that have not been attempted in any truly flexible or powerful way; and there are many linguistic phenomena, including ones as pervasive as metaphor, on which work can be hardly said to have begun. It is also the case that while appropriate forms of reasoning, like abduction, have spread from AI generally into NLP and have found useful application at more than one level of processing, there are still very intractable problems to be overcome in providing the apparatus needed to manipulate beliefs and intentions in supporting language use.

The present phase of NLP work is interesting, however, not only because of the extent to which it demonstrates that some progress has been made since the 1950s, though far less than was then expected or at least hoped for. Some of its characteristic concerns were also those of the 50s: thus as I said at the beginning, NLP has returned to some of its early themes, and by a path on an ascending spiral rather than in a closed circle, even if the ascent is slow and uneven. The present emphasis on the lexicon and on statistical information, as well as the revival of interest in MT and in retrieval, reflect the pattern illustrated, on the one hand, by Reifler's heroic efforts with the Chinese lexicon and translation (Reifler, 1967), and on the other by the earlier semantic classification work reviewed in Sparck Jones (1992). The present phase, like the first one but unlike some intervening ones, also allows for the rich idiosyncracy of language as well as for its stripped universals, and has again shifted the balance between linguistic and non-linguistic resources in language processing towards the linguistic side.

As I noted too, this return to concerns of the first phase of NLP is also a reminder of Don Walker's long-standing interests. While the Mitre work on syntax with which he was concerned (Zwicky et al., 1965) can be seen as contributing to the ample stream of computational grammar research, the concern with text data with which Don's name has been so closely associated in recent years had its foreshadowing in the title of another of his early papers: "SAFARI: an online text-processing system", a title truly symbolic for both Don and the field (Walker, 1967).

## References

- [1] ALPAC: *Language and machines: computers in translation and linguistics*, Report by the Automatic Language Processing Advisory Committee, National Academy of Science, Washington DC, 1966; see also Hutchins (1986), Chapter 8.
- [2] Alshawi, H. (ed), *The Core Language Engine*, Cambridge, MA: MIT Press, 1992.
- [3] Bledsoe, W. "I had a dream: AAAI presidential address, 19 August 1985", *The AI Magazine* 7 (1), 1986, 57-61.
- [4] Bobrow, D.G. and Collins, A. (eds) *Representation and understanding*, New York: Academic, 1975.
- [5] Booth, A.D. (ed.) *Machine translation*, Amsterdam: North-Holland, 1967.
- [6] Brady, M. and Berwick, R.C. (eds.) *Computational models of discourse*, Cambridge, MA: MIT Press, 1983.
- [7] Briscoe, E. et al. "A formalism and environment for the development of a large grammar of English", *IJCAI 87: Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 1987, 703-708.
- [8] Ceccato, S. "Correlational analysis and mechanical translation", in Booth 1967, 77-135.
- [9] Cohen, P.R., Morgan, J. and Pollack, M.E. (eds.) *Intentions in communication*, Cambridge, MA: MIT Press, 1990.
- [10] Cullingford, R. "SAM", 1981; reprinted in Grosz et al. 1986, 627-649.
- [11] Engelen, B. and McBryde, R. *Natural language markets: commercial strategies*, Ovum Ltd, 7 Rathbone Street, London, 1991.
- [12] Findler, N.V. (ed.) *Associative networks*, New York: Academic, 1979.
- [13] Galliers, J.R. and Sparck Jones, K. *Evaluating natural language processing systems*, Technical Report 291, Computer Laboratory, University of Cambridge, 1993.
- [14] Green, B.F. et al "BASEBALL: an automatic question answerer", 1961; reprinted in Grosz et al., 1986, 545-549.
- [15] Grosz, B.J., Sparck Jones, K. and Webber, B.L. (eds) *Readings in natural language processing*, Los Altos, CA: Morgan Kaufmann, 1986.
- [16] Harris, L.R. "Experience with INTELLECT", *The AI Magazine* 5(2), 1984, 43-50.
- [17] Hays, D.G. *Introduction to computational linguistics*, London: Macdonald, 1967.
- [18] Hendrix, G., Sacerdoti, E., Sagalowicz, D., Slocum, J., "Developing a Natural Language Interface to Complex Data", *ACM Transactions on Database Systems*, Vol 3, No. 3, pp 105-147, 1978.

- [19] HLT: *Proceedings of the ARPA Workshop on Human Language Technology*, March 1993; San Mateo, CA: Morgan Kaufmann, in press.
- [20] Hutchins, W.J. *Machine translation*, Chichester, England: Ellis Horwood, 1986.
- [21] Hutchins, W.J. and Somers, H.L. *An introduction to machine translation*, London: Academic Press, 1992.
- [22] Jacobs, P.S. (ed) *Text-based intelligent systems*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.
- [23] Joshi, A.K., Webber, B.L. and Sag, I.A. (eds.) *Elements of discourse understanding*, Cambridge: Cambridge University Press, 1981.
- [24] Kay, M., Gawron, J.M. and Norvig, P. *Verbmobil: a translation system for face-to-face dialogue*, CSLI, Stanford University, 1991.
- [25] Kittredge, R. and Lehrberger, J. (eds.) *Sublanguage: studies of language in restricted semantic domains*, Berlin; Walter de Gruyter, 1982.
- [26] Kobsa, A. and Wahlster, W. (eds.) *User modelling in dialogue systems*, Berlin: Springer-Verlag, 1989.
- [27] Lea, W.A. (ed) *Trends in speech recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1980.
- [28] Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages*, New York: John Wiley, 1955.
- [29] McKeown, K.R. *Text generation*, Cambridge: Cambridge University Press, 1985.
- [30] Minsky, M. (ed.) *Semantic information processing*, Cambridge, MA: MIT Press, 1968.
- [31] Minsky, M., "A framework for representing knowledge," (ed Winston, P.), *The psychology of computer vision*, McGraw-Hill, 1975.
- [32] Nagao, M. (ed) *A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, USA*, Japan Electronic Industry Development Association, 1989.
- [33] Plath, W. "Multiple path analysis and automatic translation", in Booth 1967, 267-315.
- [34] Reifler, E. "Chinese-English machine translation, its lexicographic and linguistic problems", in Booth 1967, 317-428.
- [35] Rumelhart, D.E., McClelland, J.L. and the PDP Research Group, *Parallel distributed processing*, 2 vols, Cambridge, MA: MIT Press, 1986.
- [36] Rustin, R. (ed) *Natural language processing*, New York: Algorithmics Press, 1973.

- [37] Schank, R. C., *Conceptual Information Processing*, Amsterdam, North Holland, 1975.
- [38] Schank, R.C. “Language and memory”, 1980; reprinted in Grosz et al. 1986, 171-191.
- [39] Siemens AG (ed) *Verbmobil: Mobiles Dolmetschgerat; Studie*, Siemens AG, Munchen, 1991.
- [40] Sparck Jones, K.“Natural language processing: an overview”, *International encyclopedia of linguistics* (ed W. Bright), New York: Oxford University Press, 1992, Vol. 3, 53-59.
- [41] Sparck Jones, K. “Thesaurus”, *Encyclopedia of artificial intelligence* (ed Shapiro), 2nd ed, New York: Wiley, 1992, 1605-1613.
- [42] Walker, D.E. “SAFARI: an on-line text-processing system”, *Proceedings of the American Documentation Institute Annual Meeting*, 1967, 144-147.
- [43] Winograd, T. “A procedural model of language understanding”, 1973; reprinted in Grosz et al. 1986, 249-266.
- [44] Woods, W.A. “Semantics and quantification in natural language question answering”, 1978; reprinted in Grosz et al. 1986, 205-248.
- [45] Yngve, V.H. “MT at MIT”, in Booth 1967, 451-523.
- [46] Zwicky, A.M. et al., “The MITRE syntactic analysis procedure for transformational grammars”, *Proceedings of the Fall Joint Computer Conference, 1965; AFIPS Conference Proceedings* Vol. 27, Part 1, 1965, 317-326.

# On Getting a Computer to Listen\*

Jane Robinson  
SRI International  
*e-mail: jrobinson@ai.sri.com*

Responding appropriately to matter-of-fact utterances is a common form of intelligent behavior, but specifying precisely, rather than loosely or intuitively, what takes place in even the most ordinary conversational exchanges poses formidably complex problems. The book *Understanding Spoken Language* is about those problems as they were faced in order to design and implement a computer system for processing spoken English discourse within a limited domain in an intelligent manner. "In an intelligent manner" is a key concept of the design. Because of it, the design, implementation and testing reported here are as germane to artificial intelligence, linguistics and psychology as they are to the sciences of signal processing and speech. In this preface we point out the complexities inherent in the undertaking and explain why they led us to develop components for representing and combining the kinds of knowledge people use when they understand what is said to them.

Some people object to applying words like "intelligent" to computer systems, holding that not even the license of metaphor justifies their application without evoking the connotation of "sham" or "counterfeit". Trying to define and justify the use of the word "intelligent" would be pointless. A more informative approach is to clarify our own use of it to make distinctions, leaving our readers to decide whether they would make similar distinctions using similar or different terminology.

The distinction we want to make, in the context of system design for processing continuous speech, lies somewhere along a continuum between *clever* systems and *intelligent* ones. Briefly, a merely clever system avoids the problem of assigning interpretations to indeterminate or ambiguous input by constraining the input to sound/meaning pairs that are highly differentiated in sound and completely unambiguous in meaning. A highly intelligent system, by contrast, has few or no constraints on the input other than that the utterances be those that native speakers find acceptable and understandable. It uses representations of the kinds of knowledge speakers use to constrain the problem of selecting among possible interpretations when meanings are ambiguous and sounds are difficult or even impossible to discriminate.

Here is a miniature clever system that successfully processes fourteen different utterances by displaying an abstract robot moving about, picking up an object, releasing it, and stopping. The first utterance must be among those in (1) below; the second from (2), and so on, with the option of recycling on the fourth utterance.

---

\*This article first appeared as the preface to *Understanding Spoken Language*, edited by Don Walker, which appeared in the Artificial Intelligence Series, Elsevier, North Holland, 1978. It is being reprinted here by permission of the publisher.

- |   |  |
|---|--|
| 1) a. Go on<br>b. Go back                     | c. Turn left<br>d. Turn right          |
| 2) a. Pick up a block<br>b. Pick up a pyramid | c. Grasp a block<br>d. Grasp a pyramid |
| 3) a. Release it                              | b. Drop the thing you picked up        |
| 4) a. Stop                                    | b. Return to step one                  |

What this system does, essentially, is discriminate successively between pairs of acoustic signals that are constrained to differ greatly in obvious ways. To say that it "recognizes paraphrases" or "resolves pronoun references" or "handles an infinite number of utterances" would be totally misleading.

An intelligent system, on the other hand, can accept words and phrases that sound alike but differ in meaning, because it can use other kinds of data to select correct or plausible interpretations. Given an input that cannot possibly distinguish "two went" from "to went" or "too went", it can eliminate the implausible word combinations on the basis of syntax. It can eliminate some syntactically possible combinations by using knowledge of the world. It can use representations of the domain of discourse or of past discourse history in resolving some remaining ambiguities. Such a system can also handle anaphora and reference in a general way, using mechanisms for associating pronouns and definite noun phrases with appropriate entities in its data about the world. Given the successive utterances,

- a He left his book on the table.
- b Can I read it?

the system can eliminate "table" as the coreferent of "it" even though "table" is the most recently occurring noun. Paraphrases can also be handled in a general way by building semantic representations that are the same or similar for utterances differing in vocabulary or syntax but not in meaning. An intelligent system may have mechanisms for adding and using representations of facts that can be inferred from those already stored. But rather than proceed with this description, which is after all an overview of what a reader will find in the book, we will assume that the nature of the distinction between the two systems is clear. The intelligent system can impose reasonable interpretations on indeterminate and ambiguous sounds; the clever system cannot.

Several *mixed* systems have evolved in the last few years, systems both clever and to some degree intelligent. They relax the requirements that utterances with different meanings must differ in sound, using higher level knowledge to disambiguate the underdetermined acoustic signal. However, these are not just clever systems with representations of other kinds of knowledge added on. Designing representations that make some kind of knowledge usable and providing programs for using them is itself a complex problem, but there is also the additional problem of providing for the interactions among these programs when sounds are processed in order to determine what was said. The additional layer of complexity is related to the way human perception appears to work.

It is natural to suppose that in the course of understanding what is said to us, we first hear the sounds and then interpret them. We tend to assume that perception precedes comprehension in a fairly straightforward way. But there is good evidence that our perceptions are guided by our cognitive knowledge; that our cognition focuses our attention on some parts of incoming stimuli and ignores others. In other words, once we are past the initial stages of infancy in which the world is a “blooming, buzzing confusion,” we do not simply register incoming signals in some neutral way, but we confront the scenes and sounds that impinge on us with filters for sifting out what is irrelevant to our goals and expectations of what the world is like and with frames for organizing and interpreting the meaningful remainder. And that is not all. In making our interpretations we often add to that remainder, supplying what was missing or enhancing what could only weakly be perceived.

This active processing of stimuli is perhaps easier to exemplify for vision than for hearing. When we look at a picture, we do not simply register patches of light and shade on a plane; we “see” outlines of three-dimensional objects, some of them half hidden by others. Again, this is not the end of what we do when we interpret. Suppose we are shown two pictures that are perceptibly quite different. In one, we see a couch in the foreground, partially obscuring a fireplace towards which the camera was aimed. In the other picture, there is an angled view of a fireplace with a chair to the left of it and half a couch on the right. In some objective sense, the differences in the perceptual stimuli given by the two pictures are greater than the similarities, and yet as we look at them, it dawns on us rather quickly that we are seeing the same room.

So too with hearing. When we are receiving messages, much of the incoming sound is ignored as “noise” while objectively different stimuli may sound the same to us and objectively similar ones may sound different. What impinges on our ears prompts and constrains the messages we get, but it does not determine them, and one might almost say that the sounds we perceive are determined by the messages we get.

This being so, it is not possible to program a computer to respond to relatively unconstrained speech in an intelligent manner by first processing the signal, then using acoustic-phonetic knowledge and phonological rules to classify the segments phonemically, then concatenating the segments into possible words, then parsing the words into possible sentences, and then assigning possible meanings. An intelligent system must provide for complicated interactions among its components for representing and using various kinds of knowledge. In describing our own system, the design of the mechanism for integrating the use of various kinds of knowledge has as prominent a place as descriptions of their representations themselves and the programs for applying them when they are to be used.

A final question. Do our representations and use of facts about language, discourse, and the world have analogs in human processing of what is said? The question lies in an area in which there is strong disagreement and the answer is far from clear. Certainly not everything one has to do in designing and implementing a computer system like ours provides a model or a metaphor for what people do; much is determined by the limits of the machine. This much can be said, however: designing a system for responding to speech in a nontrivial, human-like way promotes increased respect for the abilities that must underlie the most ordinary behavior people engage in.

# Utterance and Objective: Issues in Natural Language Communication

Barbara Grosz  
Harvard University  
Cambridge, MA  
*e-mail:* *grosz@das.harvard.edu*

## Abstract

Communication in natural language requires a combination of language-specific and general common-sense reasoning capabilities, the ability to represent and reason about the beliefs, goals and plans of multiple agents, and the recognition that utterances are multifaceted. This paper evaluates the capabilities of natural language processing systems against these requirements and identifies crucial areas for future research in language processing, common-sense reasoning, and their coordination. Don Walker's guiding hand can be seen in the two major premises of this paper — the importance of the communicative situation and of a consideration of language use — and in its argument for an interdisciplinary approach, as well as in the title. His strong support made the articulation of these then controversial views possible.

## 1 Introduction

Two premises, reflected in the title, underlie the perspective from which I will consider research in natural language processing in this paper.<sup>1</sup> First, progress on building computer systems that process natural languages in any meaningful sense (i.e., systems that interact reasonably with people in natural language) requires considering language as part of a larger communicative situation. In this larger situation, the participants in a conversation and their states of mind are as important to the interpretation of an utterance as the linguistic expressions from which it is formed. A central concern when language is considered as communication is its function in building and using shared models of the world. Indeed, the notion of a shared model is inherent in the word "communicate," which is derived from the Latin *communicare*, "to make common."

Second, as the phrase "utterance and objective" suggests, regarding language as communication requires consideration of what is said literally, what is intended, and the relationship between the two. Recently, the emphasis in research in natural language processing has begun to shift from an analysis of utterances as isolated linguistic phenomena to a consideration of how people use utterances to achieve certain objectives.

---

<sup>1</sup>This is a revision of a paper presented at the Sixth International Conference on Artificial Intelligence, Tokyo, Japan, August 20-24, 1979. Preparation of this paper was supported by the National Science Foundation under Grant No. MCS76-220004, and the Defense Advanced Research Projects Agency under Contract N00039-79-C0118 with the Naval Electronic Systems Command.

But, in considering objectives, it is important not to ignore the utterances themselves. A consideration of a speaker's underlying goals and motivations is critical, but so is an analysis of the particular way in which that speaker expresses his thoughts. (I will use "speaker" and "hearer" to refer respectively to the producer of an utterance and the interpreter of that utterance. Although the particular communicative environment constrains the set of linguistic and nonlinguistic devices a speaker may use (Rubin, 1977), I will ignore the differences and concentrate on those problems that are common across environments.) The choice of expression has implications for such things as what other entities may be discussed in the ensuing discourse, what the speaker's underlying beliefs (including his beliefs about the hearer) are, and what social relationship the speaker and hearer have. The reason for conjoining "utterance" and "objective" in the title of this paper is to emphasize the importance of considering both. (The similarity to *Word and Object* (Quine, 1960) is not entirely accidental. It is intended to highlight a major shift in the context in which questions about language and meaning should be considered. I believe the issues Quine raised can be addressed effectively only in this larger context.)

In the remainder of this paper I will examine three consequences of these claims for the development of language processing theories and the construction of language processing systems.

- Language processing requires a combination of language-specific mechanisms and general common-sense reasoning mechanisms. Specifying these mechanisms and their interactions constitutes a major research area.
- Because discourse involves multiple separate agents with differing conceptions of the world, language systems must be able to represent the beliefs and knowledge of multiple individual agents. The reasoning procedures that operate on these representations must be able to handle such separate beliefs. Furthermore, they must be able to operate on incomplete and sometimes inconsistent information.
- Utterances are multifaceted; they must be viewed as having effects along multiple dimensions. As a result, common-sense reasoning (especially planning) procedures must be able to handle situations that involve actions having multiple effects.

## 2 Monkeys, Bananas, and Communication

To illustrate some of the current problems in natural language processing, I will consider a variant of the "monkey and bananas" problem (McCarthy, 1968), the original version of which is substantially as follows: There is a monkey in a room in which a bunch of bananas is hanging from the ceiling, out of reach of the monkey. There is also a box in one corner of the room. The monkey's problem is to figure out what sequence of actions will get him the bananas. For a while at least, this problem was a favorite test case for automatic problems solvers, and there are several descriptions of how it can be solved by machine (e.g., see Nilsson, 1971). The variation I will discuss introduces a second monkey, the need for some communication to take place, and a change of scene to a tropical forest containing banana trees. To begin, I leave unspecified the relationship

between the two monkeys and consider the short segment of hypothetical dialogue in Illustration 1:

- (1) monkey1: I'm hungry.
- (2) monkey2: There's a stick under the old rubber tree.

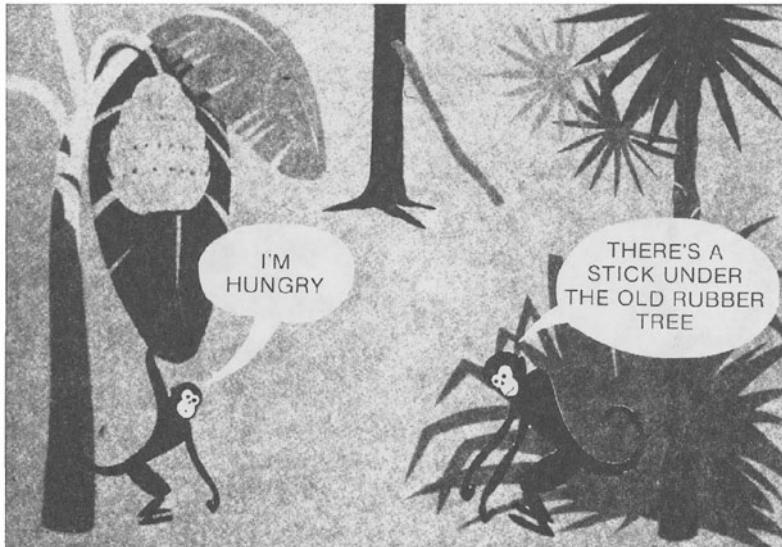


Illustration 1

If monkey1 interprets monkey2's response as most current Artificial Intelligence (AI) natural language processing systems would, he might respond with something like, "I can't eat a stick" or "I know, so what?" and, unless monkey2 helped him out, monkey1 would go hungry. Although there are a few systems now that might, with suitable tweaking, be able to get far enough for a response that indicates they have figured out that monkey2 intends for the stick to be used to knock down the bananas, there are no systems yet that would be able to understand most of the nuances of this response. For example, it implies not only that monkey2 has a plan for using the stick, but also that he expects monkey1 either to have a similar plan or to be able to figure one out once he has been told about the stick.

There is a corresponding amount of sophisticated knowledge and reasoning involved in monkey1's recognition of this request. To interpret "I'm hungry" correctly, monkey2 must recognize that a declarative statement is being used to issue a request. The robot's response in the dialogue of Illustration 2 reflects a lack of such recognition. It is inappropriate because it addresses the literal content of the monkey's statement rather than considering why he uttered it. (Notice that such a response might be appropriate in a different situation. For example, if the monkey were already eating a banana, "I'm hungry" could serve to explain why he was eating and "I understand" might serve as an

acceptance of this explanation.)

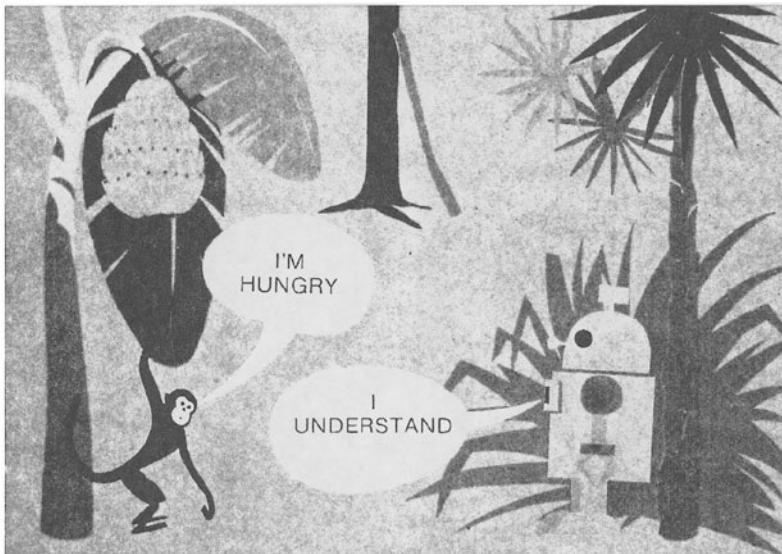


Illustration 2

A similar problem can arise with more explicit requests, like that given by the monkey in Illustration 3. Although the fact that the monkey is making a request is explicit here, the intent of his request must still be inferred. “Can you help me...” is an indirect request for assistance, not a question about the robot’s capabilities. Again the response is inappropriate because it addresses the literal content of the message rather than the intent that underlies it. Taking queries literally is a major cause of inappropriate responses by natural language processing systems (and computer systems more generally).

If we complicate the scenario just slightly, we can create a situation that would cause trouble for all current natural language processing systems. In particular, suppose that the tree the stick is under is not a rubber tree, but rather a different sort of tree. Monkey2 might still use the phrase “the rubber tree,” either by mistake or design, if he believes the phrase will suffice to enable monkey1 to identify the tree (cf. Donnellan, 1977). No current AI natural language processing system would be able to figure out where the stick is. Their responses, at best, would be like monkey1 saying, “Whaddayamean? There aren’t any rubber trees in this forest.” But referring expressions that do not accurately describe the entities they are intended to identify are typical of the sort of thing that occurs all the time in conversations between humans. The question is what will it take to get computer systems closer to being able to handle these sorts of phenomena.

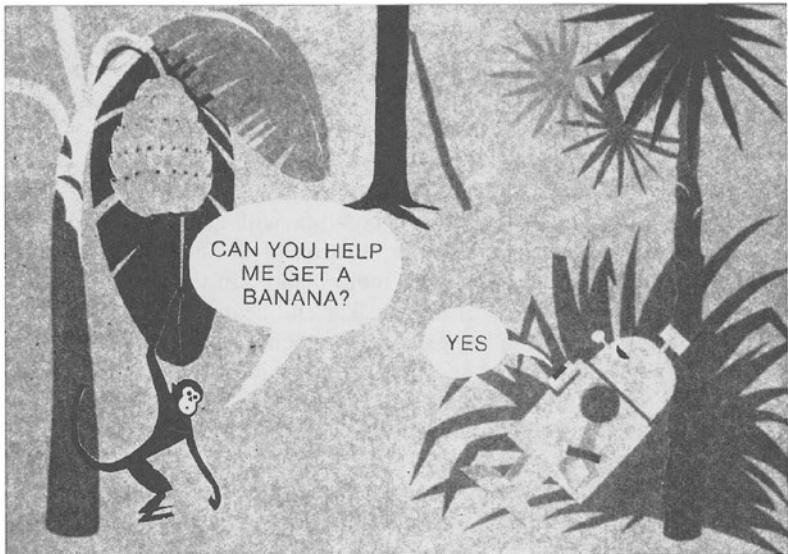


Illustration 3

In the remainder of this paper I will examine some of the research issues that need to be addressed to bring us closer to understanding why talking monkeys don't go hungry. Many of the problems that must be confronted are not confined solely to natural language processing but fall under the larger purview of AI more generally. Many critical language processing issues arise from our limited knowledge of how common-sense reasoning — which includes deduction, plausible reasoning, planning, and plan recognition — can be captured in a computational system. Consequently, research in natural language processing and research in common-sense reasoning must be tightly coordinated in the next few years.

A major source of the inadequacies of current common-sense reasoning mechanisms, when considered as possible components in a natural language processing system, is the following discrepancy. Research in problem solving and deduction has focused almost exclusively on problems that a single agent could solve alone. The need for communication arises with those problems that require the resources of multiple agents, problems that a single agent has insufficient power to solve alone. As a result, language processing is typically an issue in just those contexts where the aid of another agent is essential. To obtain that aid, the first agent must take into account the knowledge, capabilities, and goals of the second. In exchange for not needing quite as much knowledge or capability in the problem domain, the agent must have additional communication capabilities. For such problems, the option of proceeding without considering the independence of other agents and the need to communicate with them is not feasible. I believe this option is becoming less feasible as well for problem solving and deduction components used for other purposes within AI. Situations in which multiple robots must cooperate introduce similar complexity even if the communication itself can be carried out in a formal language. Sacerdoti (1978) discusses the usefulness of research in natural lan-

guage processing for the construction of distributed artificial intelligence systems. The issues being raised in this paper are central AI issues; they provide evidence of the interconnectedness of natural language processing research and other research in AI.

### 3 The Processes of Interpretation

To illustrate how language-specific processes combine with general cognitive processes (i.e., common-sense reasoning) in the interpretation of an utterance, let us consider the first monkeys and bananas example in more detail. In the following analysis, a consideration of the states of mind of the speaker and hearer will play a critical role. Each participant in a conversation brings with him a cognitive state that includes such things as a focus of attention, a set of goals to be achieved or maintained and plans for achieving them, knowledge about the domain of discourse, knowledge about how language is used, and beliefs about the cognitive states of other agents, including other participants in the current conversation. An utterance conveys information about the speaker's state; its most immediate effect is to change the hearer's state.

It is useful to view natural language interpretation as being divided into two major interacting levels. On the first, the *linguistic analysis level*, the form of an utterance is analyzed to determine its context-independent attributes. Processes at this level are concerned with determining what information is contained in the utterance itself. On the second, the *assimilation level*, common-sense reasoning processes operating in the context of the current cognitive state of the hearer use these attributes to update the cognitive state and to determine what response to the utterance is required, if any. It is important to understand that the purpose of this separation is to elucidate the kinds of processes involved in interpretation. The actual flow of processing during interpretation entails a great deal of interaction among the processes in the different levels, and there are major research issues concerned with their coordination (e.g., see Robinson, 1980a; Walker, 1978).

To illustrate these levels, let us return to the example and consider the interpretation of monkey2's response (2), "There's a stick under the old rubber tree," to monkey1's indirect request (1).

#### 3.1 Linguistic Analysis

At this level, the parsing process that assigns syntactic structure to the utterance also assigns attributes to the various syntactic subphrases in the utterance and to the utterance as a whole. Many of these attributes are of a semantic or pragmatic nature. For example, the attributes of the phrase "the old rubber tree" might include

- The phrase is of syntactic class NP (noun phrase).
- The phrase is definitely determined.
- The phrase describes a  $t$  such that TREE( $t$ ) and OLD( $t, T$ ), where OLD and TREE are predicate symbols and the second argument to the predicate OLD indicates the set with respect to which age is evaluated.

I have left open the question of what happens with the modifier “rubber”; suffice it to say, the question of how it modifies cannot be resolved solely at the linguistic level. In general, the question of how much semantic specificity should be imposed at the linguistic level is an open research question.

Attributes of a complete utterance include such properties as its syntactic structure and the presuppositions (or, implicit assumptions) and assertions it makes. (Although what an utterance presupposes and asserts are not necessarily components of the intended meaning, the recognition of presuppositions and assertions is prerequisite to the assimilation level of processing.) Attributes of utterance (2) as a whole include:

- The utterance presupposes that there exists a  $t$  such that  $\text{OLD}(t, T)$  and  $\text{TREE}(t)$ , and that the description “ $\text{OLD}(t, T) \& \text{TREE}(t)$ ” should allow  $t$  to be determined uniquely in the current context.
- The utterance asserts that there exists an  $s$  such that  $\text{STICK}(s)$ .
- The utterance asserts that  $\text{UNDER}(s, t)$ .

### 3.2 Assimilation

As attributes are extracted through the parsing process at the linguistic analysis level, common-sense reasoning processes begin to act on those attributes at the assimilation level. Two major activities are involved: completing the literal interpretation of an utterance in context, and drawing implications from that interpretation to discover the intended meaning.

For the example utterance (2), completing the literal interpretation in context involves the identification of the referent of the definite noun phrase, “the old rubber tree.” The first attribute above indicates that a unique tree should be easily identified in context. Those objects currently in monkey1’s focus of attention are examined (perhaps requiring sophisticated common-sense reasoning) to determine whether there is such a tree among them. Assume that none is found. It may be that only two kinds of trees are present in this forest, and that one kind, say gumgum trees, resemble rubber trees, and that of all the trees near the two monkeys only one is a gumgum tree. Monkey1 may tentatively assume that “rubber tree” matches “gumgum tree” closely enough to serve to identify this tree.

The sentence says there’s a stick under the tree, so monkey1 might look under the tree and discover that, indeed, there is exactly one stick there. That stick must be the stick whose existence monkey2 was informing him of. The literal interpretation of the utterance is seen to be that the newly found stick is under the gumgum tree. (For more complex utterances, the process of completing the literal interpretation can involve determining the scopes of quantifiers and resolving various types of ambiguities.)

Knowing that the sentence presupposed the existence of a rubber tree and asserted the existence of a stick, monkey1 may infer that monkey2 believes these presuppositions. Thus, monkey1 comes to believe several new things about monkey2’s beliefs; in particular, that he believes these two entities exist, and that he thinks the gumgum tree is a rubber tree, or at least thinks that this description can be used to identify the tree. This fact may be important in further communications. Monkey1 may also infer

that because monkey2 has just mentioned the stick and the tree, they are in his focus of attention and that he (monkey1), too, should pay special attention to these objects. The stick may be of particular importance because it was the subject of a there-insertion sentence (a syntactic position of prominence) and has been newly introduced into his focus of attention.

The second major process of assimilation is to use common-sense reasoning to determine how the utterance fits into the current set of plans and goals. In general, this is a highly complex process.<sup>2</sup> For the particular example of interpreting utterance (2) in the context implied by utterance (1), monkey1 must determine that, "There's a stick under the rubber tree," has to do with his problem of getting something to eat. Briefly, he must see that the sentence emphasizes the stick and must know (or infer) that such sticks are often useful tools for getting things out of trees. He must infer that monkey2 intends for him to use this stick in conjunction with a standard plan for knocking down things to acquire some bananas and accomplish his (implicitly stated) goal of not being hungry.

## 4 The Multifaceted Nature of Utterances

Just as an agent may perform physical actions intended to alter the physical state of his environment, he may perform linguistic actions (utter sentences) intended to alter the cognitive state of the hearer. To determine what objectives an utterance is intended to achieve requires determining where that utterance fits in the speaker's plans. But because a single utterance may be used to achieve multiple effects simultaneously, the problem is more complex than either the analogy with physical actions or the preceding examples at first seem to suggest. (Physical actions may also have effects along multiple dimensions although they are not usually thought of as doing so. For example, the action of slamming a door in someone's face not only results in the door being closed, but also communicates anger.)

The discussion so far has concentrated on a single dimension of effect: the use of an utterance to achieve what I will call a domain goal, that is, to convey information about the domain of discourse. In this section I want to discuss two other dimensions along which an utterance can have effects — the social and the discourse — and look at some of the problems in interpretation and generation that arise from the multifaceted nature of utterances.<sup>3</sup>

To illustrate the three dimensions, consider the following utterance made by the hungry monkey in our illustrations (in this instance assume he sees the stick and realizes it can be used to knock down some bananas),

"Please hand me the stick."

At the domain level, the utterance expresses a proposition that might be written as HAND (MONKEY2, MONKEY1, S1), where MONKEY1 refers to monkey1 (the

---

<sup>2</sup>The complexity is well illustrated by the analysis of a set of therapeutic interviews in Labov and Fanshel, (1977).

<sup>3</sup>These dimensions parallel the three functions of language – ideational, interpersonal, and textual – in Halliday (1970), but the perspective I take on them is closer to that presented in Levy (1978).

hungry monkey), MONKEY2 to monkey2, S1 to the stick under the tree, and HAND to the operation of transferring some object (given in the third argument) from one agent (first argument) to another (second argument) by hand. General domain information such as the taxonomic relationship that HAND is a kind of GIVE, and plan-based information about using the stick are an implicit part of the interpretation of the utterance along this dimension. At the social level, the utterance is a request; its imperative mood is modified by “please.” At the discourse level, the utterance identifies and focuses on the stick S1.

The social dimension includes those aspects of an utterance that concern the establishment and maintenance of interpersonal relationships. This dimension of utterance (1), “I’m hungry,” is easily seen when it is compared with such choices as

- (3) “How can I get some of those blasted bananas?”
- (4) “Can you help me get some bananas down?”
- (5) “Get me a banana.”

Each of these achieves the same domain goal, informing monkey2 of monkey1’s desire to obtain some bananas, but utterance (1) does not convey the same familiarity as utterance (3) or the same level of frustration. (The bananas, after all, are not “blasted.”) Similarly, utterance (4) makes the same request as utterance (5) but does so indirectly. A big monkey might use (5) to a small monkey and get a banana, but if a small monkey uttered it to a big monkey, he would more likely get a response like, “Not until you show some respect for your elders.” A typical use of indirect speech acts like (4) is to moderate requests.

The social dimension is present in every discourse<sup>4</sup> and prevails in some (e.g., Hobbs, 1979). It has been largely ignored in natural language processing research to date. However, any analysis that translates the utterances in (1) and (3)-(5) only into requests for help getting food, misses a significant part of the meaning of each of the utterances. An assumption has been that some sort of neutral stance is possible. But, even the choice of the unmarked (neutral) case is a choice; not choosing is choosing not to choose (cf. Goffman, 1978). Although there are some serious philosophical issues raised by this dimension of utterances when considering communication between people and computers, I do not think we can continue to ignore it.

The *discourse dimension* includes those aspects of an utterance that derive from its participation in a coherent discourse — how the utterance relates to the utterances that preceded it and to what will follow. Although language is linear (only one word can be uttered at a time), the information a speaker has to convey typically is highly interconnected. As a result, the speaker must use multiple utterances to convey it. Each individual utterance must contain information that provides links to what went before and properly set the stage for what follows. Utterances that convey the same propositional content may differ widely in such things as the entities they indicate a speaker is focused on and hence may refer to later. As an extreme example, note that the propositional content of “Not every stick isn’t under the rubber tree” is equivalent to that of utterance

---

<sup>4</sup>Pittenger et al. (1960) point out that “no matter what else human beings may be communicating about, or may think they are communicating about, *they are always communicating about themselves, about one another, and about the immediate context of the communication.*”

(2), but because it does not mention any individual stick, it does not allow whoever speaks next to make any reference to the stick that is under the gumgum tree.<sup>5</sup>

There are two characteristics of these dimensions and the multifaceted nature of utterances that introduce complications into natural language processing. First, as Halliday (1977) has pointed out, the units in which the information is conveyed along these other dimensions of meaning do not follow the constituent structure of sentences nearly so nicely as do the units conveying propositional content. In particular, the social implications of an utterance are typically reflected in choices scattered throughout it; for example, they are reflected in the choice of utterance type (a request vs. a command) and in the choice of lexical items.

Second, an utterance may relate to plans and goals along any number of these dimensions. It may be a comment on the preceding utterance itself, its social implications (or both, as is usually the case with "I shouldn't have said that"), or on some part of the domain content of the utterance. It is not simply a matter of determining where an utterance fits into a speaker's plan, but of determining which plan or plans — domain, social, or discourse — the utterance fits into. A one dimensional analysis of an utterance is insufficient to capture the different effects (cf. Goffman, 1978).

The multifaceted nature of utterances poses problems for language generation as well. A speaker typically must coordinate goals along each of these dimensions. He must design an utterance that conveys information linking it to the preceding discourse and maintains the social relationship he has with the hearer(s) (or establishes one) as well as conveying domain-specific information.<sup>6</sup> The speaker's task is further complicated because he has only incomplete knowledge of the intended hearer's goals, plans and beliefs.

## 5 State of Art

I will use our work in natural language processing at SRI International (Robinson, 1980a; Walker, 1978) as an exemplar for discussing the current state of research in this area, both because I am most familiar with it and because I think the framework it provides is a useful one for seeing not only where the field stands, but also where the next several years effort might best be expended. A caveat is necessary before proceeding. The discussion that follows considers only research concerned with developing theoretical models of language use and the systems that contribute to this research. Because of space limitations, I will not discuss a second major direction of current research in natural language processing, that concerned with the construction of practical natural language interfaces (e.g., Hendrix et al., 1978). The major difference between the two kinds of efforts is that research on interfaces has separated language processing from the rest of the system whereas one of the major concerns of research in the more theoretical direction is the interaction between language-specific and general knowledge and reasoning in the context of communication.

---

<sup>5</sup>This example is based on one suggested by Barbara Partee for the Sloan Workshop at the University of Massachusetts, December, 1978. A discussion of her example is included in Grosz and Hendrix, 1978.

<sup>6</sup>Levy discusses how the multiple levels along which a speaker plans are reflected in what he says and the structure of his discourse.

SRI's TDUS system has been constructed as part of a research effort directed at investigating the knowledge and processes needed for participation in task-oriented dialogues (Robinson, 1980a). The system participates in a dialogue with a user about the performance of an assembly task. It coordinates multiple sources of language-specific knowledge and combines them with certain general knowledge and common-sense reasoning strategies in arriving at a literal interpretation of an utterance in the context of an ongoing task-oriented dialogue.<sup>7</sup> A major feature of the system is the tight coupling of syntactic form and semantic interpretation. In the interpretation of an utterance, it associates collections of attributes with each phrase. For example, noun phrases are annotated with values for the attribute "definiteness," a property that is relevant for drawing inferences about focusing (Grosz, 1977a, 1977b, 1980) and about presuppositions of existence and mutual knowledge (Clark and Marshall, 1980).

Interpretation is performed in multiple stages under control of an executive and in accordance with the specifications of a language definition that coordinates multiple "knowledge sources" for interpreting each phrase. Two sorts of processes take part in the linguistic level of analysis. First, there are processes that interpret the input "bottom up" (i.e., words  $\Rightarrow$  phrases  $\Rightarrow$  larger phrases  $\Rightarrow$  sentences). In the analysis of utterance (2), these processes would provide attributes specifying that the phrase "a stick" is indefinite and in the subject position of a there-initial sentence. They would specify that the phrase "the rubber tree" is definite and presupposes the existence of a uniquely identifiable entity. Second, there are processes that refine the interpretation of a phrase in the context of the larger phrases that contain it, doing such things as establishing a relationship between syntactic units and *descriptions* of (sets of propositions about) objects in the domain model. For example, the structure for "the rubber tree" would include formal logical expressions regarding existence and treeness.

The assimilation level in the current system only goes so far as determining a literal interpretation in context. The major tasks performed here include delimiting the scope of quantifiers and associating references to objects with particular entities in the domain model, taking into account the overall dialogue and task context. To perform these tasks, the system includes mechanisms for representing and reasoning about complex processes (Appelt et al., 1980). In the case of our two monkeys, the system would determine whether there was a unique rubber tree in, or near, the focus of attention of the monkey (more on this shortly) and then posit, or check, the existence of a stick under it.

Although it only interprets utterances literally, TDUS does make some inferences based on the information explicitly contained in an utterance. The plans it knows about are partially ordered (and not linear), and the structures it uses allow for describing plans at multiple levels of abstraction. To see the sorts of inferences TDUS will make, consider the sequence:

**(6) User: I am attaching the pump.**

---

<sup>7</sup>Several other systems are capable of fairly sophisticated analysis and processing at the level of coordinating different kinds of language-specific capabilities (e.g., Sager and Grishman, 1975; Landsbergen, 1976; Plath, 1976; Woods et al., 1976; Bobrow et al., 1977; Reddy et al., 1977) and of taking into account some of the ways in which context affects meaning through the application of limited action scenarios (Schank et al., 1975; Novak, 1977) or by considering (either independently or in conjunction with such scenarios) language-specific mechanisms that reference context (Hobbs, 1976; Rieger, 1975; Hayes, 1978; Mann et al., 1977; Sidner, 1979).

(7) System: OK

(8) User: Which wrench should I use to bolt it?

In interpreting utterance (6), the system updates its model of the task of attaching the pump. It uses tense and aspect information to determine that the task has been started but not completed (the user said, "am attaching," not "have attached"). As part of interpreting this utterance, the system also records that the user is now focusing on the pump and the attaching operation. The system uses this focusing information and information in its model of the task to determine that the bolting operation referred to is a substep of the attaching operation and that the "it" in utterance (8) is being used to refer to the pump. In addition, TDUS infers that all of the substeps of the attaching operation that had to precede the bolting have been done (Appelt et al., 1980; Robinson, 1980b).

Initial progress has been made in overcoming the limitations of literal interpretation and including a consideration of a speaker's plans and goals in the interpretation of an utterance. Recent research on the role of planning in language processing includes that of Cohen (1978), Wilensky (1978), Carbonell (1979) and Allen (1979). Cohen (1978) views speech acts (Searle, 1969) as one kind of goal-oriented activity and describes a system that uses mechanisms previously used for planning nonlinguistic actions to plan individual speech acts (on the level of requesting and informing) intended to satisfy some goals involving the speaker's or hearer's knowledge. In Wilensky's work on story understanding (see also Schank and Abelson, 1977), the speaker's overall plans and goals, some of which are implicit, are inferred from substeps and intermediate or triggering states (e.g., inferring from "John was hungry. He got in his car." that John was going to get something to eat). Carbonell (1979) describes a system constructed to investigate how two agents with different goals interpret an input differently; it is particularly concerned with the effect of conflicting plans on interpretation. Allen (1979) describes a system based on a model in which speech acts are defined in terms of "the plan the hearer believes the speaker intended him to recognize" and has perhaps gone furthest in determining mechanisms by which a speaker's goals and plans can be taken into account in the interpretation of an utterance.

These efforts have demonstrated the feasibility of incorporating planning and plan recognition into the common-sense reasoning component of a natural language processing system, but their limitations highlight the need for more robust capabilities in order to achieve the integration of language-specific and general common-sense reasoning capabilities required for fluent communication in natural language. No system combines a consideration of multiple agents having different goals with a consideration of the problems that arise from multiple agents having separate beliefs and each having only incomplete knowledge about the other agent's plans and goals.<sup>8</sup> Furthermore, only simple sequences of actions have been considered, and no attempt has been made to treat hypothetical worlds.

One of the major weaknesses in current AI systems and theories (and the limitation of current systems that I find of most concern) is that they consider utterances as having a single meaning or effect. Analogously, a critical omission in work on planning and

---

<sup>8</sup>Moore (1979) discusses problems of reasoning about knowledge and belief.

language is that it fails to consider the multiple dimensions on which an utterance can have effects. If utterances are considered operators (where “operator” is meant in the general sense of something that produces an effect), they must be viewed as *conglomerate* operators.

Although it does not yet go beyond literal interpretation (except by filling in unmentioned intermediate steps in the task being performed), TDUS does account for two kinds of effects of an utterance. In addition to determining the propositional content of an utterance (and what it literally conveys about the state of the world), the system determines whether the utterance indicates that the speaker’s focus of attention has shifted (Grosz 1977a, 1977b, 1980; Sidner 1979).<sup>9</sup>

To summarize then, one or more of the following crucial limitations is evident in every natural language processing system constructed to date (although most of these problems have been addressed to some extent in the research described above and elsewhere):

- Interpretation is literal (only propositional content is determined).
- The knowledge and beliefs of all participants in a discourse are assumed to be identical.
- The plans and goals of all participants are considered to be identical.
- The multifaceted nature of utterances is not considered.

To move beyond this state, the major problems to be faced at the level of linguistic analysis concern determining how different linguistic constructions are used to convey information about such things as the speaker’s (implicit) assumptions about the hearer’s beliefs, what entities the speaker is focusing on, and the speaker’s attitude toward the hearer. The problems to be faced at the assimilation level are more fundamental. In particular, we need to determine common-sense reasoning mechanisms that can derive complex connections between plans and goals — connections that are not explicit either in the dialogue or in the plans and goals themselves — and to reason about these relationships in an environment where the problem solver’s knowledge is necessarily incomplete. This is not just a matter of specifying more details of particular relationships, but of specifying new kinds of problem solving and reasoning structures and procedures that operate in the kind of environment in which natural language communication usually occurs.

## 6 Common-Sense Reasoning in Natural Language Processing

The previous sections of this paper have suggested several complexities in the common-sense reasoning needs of natural language communication. A participant in a communicative situation typically has incomplete information about other participants. In

---

<sup>9</sup>Grosz and Hendrix (1978) discuss focusing as one of the segments of cognitive state crucial to the interpretation of both definite and indefinite referring expressions, and Grosz (1980) discusses several open problems in modeling the focusing process.

particular he cannot assume that their beliefs, goals, or plans are identical. Communication is inherently interpersonal. Furthermore, the information a speaker conveys typically requires a sequence of utterances. As a result, interpretation requires recognition of different kinds of plans, and generation requires the ability to coordinate multiple kinds of actions to satisfy goals along multiple dimensions. Other complications are introduced by the interactions among plans of different agents (Bruce and Newman, 1978; Hobbs and Robinson (1978) discuss some of the complexity of the relationship between an utterance and domain specific plans.)

From this perspective, the current deduction and planning systems in AI are deficient in several areas critical for natural language processing. A review of the current state of the art in plan generation and recognition shows that the most advanced systems have one or another (but not both) of the following capabilities: plans for partially ordered sequences of actions can be generated (Sacerdoti, 1977) and recognized (Genesereth, 1978; Schmidt and Sridharan, 1977) at multiple levels of detail in a restricted subject area. However, these programs only consider single agents, assume the system's view of the world is "the correct" one, and plan for actions that produce a state change characterized by a single primary effect.

The most important directions in which these capabilities must be extended and integrated for use in the interpretation and generation of language are the following:

- It must be possible to plan in a dynamic environment that includes other active agents, given incomplete information.
- It must be possible to coordinate different types of actions and plan to achieve multiple primary effects simultaneously.
- It must be possible to recognize previously unanticipated plans.

## 7 Conclusions

Common-sense reasoning, especially planning, is a central issue in language research, not only within artificial intelligence, but also in linguistics (e.g., Chafe, 1978; Morgan, 1978), sociolinguistics (e.g., Kasher, 1978). The literal content of an utterance must be interpreted within the context of the beliefs, goals, and plans of the dialogue participants, so that a hearer can move beyond literal content to the intentions that lie behind the utterance. Furthermore, it is insufficient to consider an utterance as being addressed to a single purpose. Typically, an utterance serves multiple purposes: it highlights certain objects and relationships, conveys an attitude toward them, and provides links to previous utterances in addition to communicating some propositional content.

Progress toward understanding the relationship between utterances and objectives and its effect on natural language communication will be best furthered by consideration of the fundamental linguistic, common-sense reasoning, and planning processes involved in language use and their interaction. A merger of research in common-sense reasoning and language processing is an important goal both for developing a computational theory of the communicative use of language and for constructing computer-based natural language processing systems. The next few years of research on language processing

should be concerned to a large extent with issues that are at least as much issues of common-sense reasoning (especially planning issues). While common-sense reasoning research could continue without any regard for language, there is some evidence that the perspective of language processing will provide insights into fundamental issues in planning that confront AI more generally.

Finally, I want to emphasize the long-term nature of the problems that confront natural language processing research in AI. I believe we should start by adding communication capabilities to systems that have solid capabilities in solving some problem (constructing such systems first if necessary; cf. McDermott, 1976). Although it may initially take longer to create functioning systems, the systems that result will be useful, not toys. People will have a reason to communicate with such systems. Monkey2 can help monkey1 get something to eat only if he himself has a realistic conception of the complexities of monkey1's world.

### Acknowledgments

The Natural Language Research Group of the Artificial Intelligence Center at SRI International provides a stimulating environment in which to pursue research in language processing. Many of the ideas presented in this paper are the product of thought-provoking discussions, both oral and written, with members of this group, especially Gary Hendrix, Douglas Appelt, Jerry Hobbs, Robert Moore, Nils Nilsson, Ann Robinson, Jane Robinson, Earl Sacerdoti, and Donald Walker. I would also like to thank David Levy and Mitch Marcus for many discussions about language and the insights they provided. Many of these people also provided helpful comments on this paper.

## References

- [1] Allen, J. "A Plan-Based Approach to Speech Act Recognition," Technical Report No. 131/79, Department of Computer Science, University of Toronto, 1979.
- [2] Appelt, D., B. Grosz, G. Hendrix, A. Robinson, "The representation and use of process knowledge," Technical Note 207, Artificial Intelligence Center, SRI International, Menlo Park, CA, 1980.
- [3] Bobrow, D. et al., "GUS, A Frame Driven Dialog System," *Artificial Intelligence*, 8, 1977, 155-173.
- [4] Bruce, B., D. Newman, "Interacting Plans," *Cognitive Science*, 2, 1978, 195-233.
- [5] Carbonell, J., "Computer Models of Social and Political Reasoning," Ph. D. Thesis, Yale University, New Haven, CT, 1979.
- [6] Chafe, W., "The Flow of Thought and the Flow of Language," in T. Givón (ed.), *Syntax and Semantics*, 12, Academic Press, New York, 1978.

- [7] Clark, H., C. Marshall, "Definite Reference and Mutual Knowledge," in A. Joshi, I. Sag, B. Webber (eds.), *Elements of Discourse Understanding: Proceedings of a Workshop on Computational Aspects of Linguistic Structure and Discourse Setting*, Cambridge University Press, Cambridge, England, 1980.
- [8] Cohen, P., "On Knowing What to Say: Planning Speech Acts," Technical Report No. 118, Department of Computer Science, University of Toronto, Toronto, Canada, 1978.
- [9] Collins, A., "Studies of Plausible Reasoning," Volume I, BBN Report No. 3810, Bolt Beranek and Newman, Cambridge, MA, 1978.
- [10] Donnellan, K., "Reference and Definite Descriptions," in S. Schwartz (ed.), *Naming, Necessity and Natural Kind*, Cornell University Press, Ithaca, NY, 1977, 42-65.
- [11] Genesereth, M., "Automated Consultation for Complex Computer Systems," Ph.D. Thesis, Harvard University, Cambridge, MA, 1978.
- [12] Goffman, E., *Frame analysis*, Harper, New York, 1974.
- [13] Goffman, E., "Response Cries," *Language*, Vol. 54, 1978, 787-815.
- [14] Grosz, B., "The Representation and Use of Focus in Dialogue Understanding," Technical Note 151, Artificial Intelligence Center, SRI International, Menlo Park, CA, 1977a.
- [15] Grosz, B., "The Representation and Use of Focus in a System for Understanding Dialogs," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Carnegie-Mellon University, Pittsburgh, PA, 1977b, 67-76.
- [16] Grosz, B., "Focusing and description in natural language dialogues," in A. Joshi, I. Sag, B. Webber (eds.), *Elements of Discourse Understanding: Proceedings of a Workshop on Computational Aspects of Linguistic Structure and Discourse Setting*, Cambridge University Press, Cambridge, England, 1980.
- [17] Grosz, B., G. Hendrix, "A Computational Perspective on Indefinite Reference," presented at Sloan Workshop on Indefinite Reference, University of Massachusetts, Amherst, MA, December 1978. Also Technical Note No. 181, Artificial Intelligence Center, SRI International, Menlo Park, CA.
- [18] Halliday, M., "Language Structure and Language Function," in J. Lyons (ed.), *New Horizons in Linguistics*, Penguin, 1970.
- [19] Halliday, M., "Structure and Function in Language," presented at Symposium on Discourse and Syntax, University of California at Los Angeles, Los Angeles, CA, November 1977.
- [20] Hayes, P., "Some Association-Based Techniques for Lexical Disambiguation by Machine," Doctoral Thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 1978.

- [21] Hendrix, G., E. Sacerdoti, D. Sagalowicz, J. Slocum, "Developing a Natural Language Interface to Complex Data," *ACM Transactions on Database Systems*, Vol. 3, No. 2, June 1978, 105-147.
- [22] Hobbs, J., "A Computational Approach to Discourse Analysis," Research Report 76-2, Department of Computer Sciences, City College, City University of New York, New York, NY, 1976.
- [23] Hobbs, J., "Conversation as Planned Behavior," in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, Stanford University, Stanford, CA, 1979.
- [24] Hobbs, J., J. Robinson, "Why Ask?" SRI Technical Note 169, SRI International, Menlo Park, CA, 1978.
- [25] Kasher, A., "Indefinite Reference: Indispensability of Pragmatics," presented at Sloan Workshop on Indefinite Reference, University of Massachusetts, Amherst, MA, December 1978.
- [26] Labov, W., D. Fanshel, *Therapeutic Discourse*, Academic Press, New York, 1977.
- [27] Landsbergen, S., "Syntax and Formal Semantics of English in PHLIQAI," in *Coling 76, Preprints of the 6th International Conference on Computational Linguistics*, Ottawa, Ontario, Canada, June 28–July 2, 1976, No. 21.
- [28] Levy, D., "Communicative Goals and Strategies: Between Discourse and Syntax," in T. Givon (ed.), *Syntax and Semantics*, 12, Academic Press, New York, 1978.
- [29] Mann, W., J. Moore, J. Levin, "A comprehension model for human dialogue," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Carnegie-Mellon University, Pittsburgh, PA, 1977, 77-87.
- [30] McCarthy, J., "Programs with Common Sense," in M. Minsky (ed.), *Semantic Information Processing*, MIT Press, Cambridge, MA, 1968, 403-419.
- [31] McDermott, D., "Artificial Intelligence Meets Natural Stupidity," *SIGART Newsletter*, no. 57, 1976, 4-9.
- [32] Moore, R., "Reasoning about Knowledge and Actions," Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1979.
- [33] Morgan, J., "Two Types of Convention in Indirect Speech Acts", in P. Cole (ed.), *Syntax and Semantics*, Vol. 9, Academic Press, New York, 1978.
- [34] Nilsson, N., *Problem Solving Methods in Artificial Intelligence*, McGraw-Hill, New York, 1971.
- [35] Novak, G., "Representations of knowledge in a program for solving physics problems," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Carnegie-Mellon University, Pittsburgh, PA, 1977, 286-291.

- [36] Pittenger, R., C. Hockett, J. Danehy, *The First Five Minutes*, Paul Martineau, Ithaca, NY, 1960.
- [37] Plath, W., "Request: A Natural Language Question-Answering System," *IBM Journal of Research and Development*, 20, 4, 1976, 326-335.
- [38] Quine, W., *Word and Object*, MIT Press, Cambridge, MA, 1960.
- [39] Reddy, D. et al, "Speech Understanding Systems: A Summary of Results of the Five-Year Research Effort," Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1977.
- [40] Rieger, C., "Conceptual Overlays: A Mechanism for the Interpretation of Sentence Meaning in Context," Technical Report TR-354, Computer Science Department, University of Maryland, College Park, MD, 1975.
- [41] Robinson, A., "Understanding Natural-Language Utterances in Dialogs about Tasks," Technical Note 210, SRI International, Menlo Park, CA, 1980a.
- [42] Robinson, A., "Interpreting Verb Phrase References in Dialogs," in *Proceedings of the Third Conference of The Canadian Society for Computational Studies of Intelligence*, Victoria, British Columbia, May 14-16, 1980b.
- [43] Rubin, A., "A Theoretical Taxonomy of the Differences Between Oral and Written Language," in R. Spiro et al. (eds.), *Theoretical Issues in Reading Comprehension*, Lawrence Erlbaum, Hillsdale, NJ, 1978.
- [44] Sacerdoti, E., *A Structure for Plans and Behavior*, Elsevier, NY, 1977.
- [45] Sacerdoti, E., "What Language Understanding Research Suggests about Distributed Artificial Intelligence," Proceedings of a workshop held at Carnegie-Mellon University, December 7-8, 1978.
- [46] Sager, N., R. Grishman, "The Restriction Language for Computer Grammars," *Communications of the ACM*, 18, 1975, 390-400.
- [47] Schank, R., and Yale A.I. Project 1975, "SAM-A story understander," Research Report No. 43, Department of Computer Science, Yale University, New Haven, CT, 1975.
- [48] Schank, R., R. Abelson, *Scripts, Plans, Goals, and Understanding*, Laurence Erlbaum Associates, Hillsdale, NJ, 1977.
- [49] Schmidt, C., N.S. Sridharan, "Plan Recognition Using a Hypothesis and Revise Paradigm: An Example," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Carnegie-Mellon University, Pittsburgh, PA, 1977, 480-486.
- [50] Searle, J., *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge, England, 1969.

- [51] Sidner, C., “A Computational Model of Co-Reference Comprehension in English,” Ph. D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1979.
- [52] Walker, D. (ed.), *Understanding Spoken Language*, Elsevier North-Holland, NY, 1978.
- [53] Wilensky, R., “Understanding Goal-Based Stories,” Ph.D. Thesis, Yale University, New Haven, CT, 1978.
- [54] Woods, W. et al., “Speech Understanding Systems: Final Report,” BBN Report No. 3438, Bolt Beranek and Newman, Cambridge, MA, 1976.

# On the Proper Place of Semantics in Machine Translation\*

Margaret King  
ISSCO  
University of Geneva  
*e-mail: king@divsun.unige.ch*

## Preface

This paper owes much, both directly and indirectly, to Don Walker. Indirectly, it concerns a topic which was of great interest to both of us, and which we often discussed. Don helped me to clarify my own ideas and was always both perceptive and practical, a rare combination which made discussion with him constantly illuminating. There is also a direct debt; it was written at a time when I was hors combat from the normal round of life. Don encouraged me to use the opportunity to work out what I really thought without the usual time pressure, and cheered me up in the patches of loneliness. In brief, he lent me some of his own courage.

I am very grateful to Professor Makato Nagao and to the original publishers for allowing me to contribute something which brings back to me strong and good memories of a friendship I valued enormously.

## 1 Introduction

In this paper I want to ask first what it means to be able to translate between two languages, then use the results to examine the status of a number of semantics based mechanisms frequently proposed for use in machine translation systems. The thrust of the argument will be that it makes no sense to look for semantic or epistemological universals, with the intention of basing translation on language independent abstract entities, but that such mechanisms as semantic features or deep case roles should rather be thought of as engineering tools to be used within a view of translation as essentially a linguistic enterprise.

---

\*This paper is reprinted with permission from M. Nagao, ed., *Language and Artificial Intelligence*, North Holland, Elsevier Science Publishers B.V., 1987

## 2 Meaning and Translation

Much of what one believes about translation depends on what one believes about meaning. A once prevalent view of the question in philosophy would have us believe that somehow the ‘meanings of words are entities independent of the words themselves’, so that for any word or phrase, there is something which could be called the meaning of that word or phrase. Thus, a theory of meaning would consist in trying to say in a general way what kinds of things meanings were, with there perhaps being several different kinds of meanings, depending on the theory. Quine gives a graphic summary of this view at the beginning of *The inscrutability of reference*: ‘Uncritical semantics is the myth of a museum in which the exhibits are meanings and the words are labels. To switch languages is to change the labels.’

I shall not, here argue against the ‘independent entities’ view in any detail: that has been done by many others before, and far better than I could do it. (For example, by Wittgenstein, Ryle, Austin and Quine himself - to name but a few).

What is more interesting for us here is to look at some of the alternatives that have been put forward, and see what the consequences are for a view on translation. Nearly all the alternatives start by denying the assumption that one can name or identify the meaning of a phrase or a word, and by suggesting that we have been misled by the fact that one can ask ‘What does x mean?’ into believing that that question is the same sort of question as ‘Who is Rachel’s maths teacher?’, which clearly has a coherent answer by naming ('Mr. Dupont') or by identifying ('the man with red hair'). This sounds rather like just saying that the view is wrong and going no further, but if we look at the way we normally talk about meaning, for example, at what is involved in explaining the meaning of a word and at what is pre-supposed by our ways of explaining, it becomes rather clear that we do not appeal to any notion of independent meanings, and thus the charge that a wrong question has been asked gains some support. Both Wittgenstein and Austin approach the question in this way, by asking how we explain what something means. This leads to the suggestion that we substitute for the ‘independent meanings’ view, the ordinary conception of meaning, what Caton calls ‘the everyday concept in daily use.’ The most famous statement of this is Wittgenstein’s, in the *Philosophical Investigations* ‘For a large class of cases - though not for all - in which we employ the word “meaning” it can be defined thus: the meaning of a word is its use in the language. (1953, 34).

Another way of stating this would be to say that there is no point in looking for the meaning of a word or phrase outside the language in which it is a word or a phrase: in other words, meaning is language internal and critically bound to the particular language. (Actually, it is not totally clear that Wittgenstein himself would accept this reformulation: he may, in the passage cited above, simply have been recommending us to look at how ‘meaning’ is used in ordinary language, in the same way that Austin was ‘explaining the syntaxics’ and ‘demonstrating the semantics’ of ‘meaning’. It is however very close to the view put forward in Ryle’s *Ordinary Language*, 1953).

If meaning is language internal, what can one then say about the relation between language and the external world? The external world is clearly there: chairs can be sat on and windows broken. But it would seem to follow from this view that language imposes structure on the world and not vice versa, and that a particular language will

determine a particular structuring of the world.

An alternative way of approaching the question, that followed by Quine, will, in the end, lead to the same conclusion. The ‘independent entities’ view rather naturally leads to a pre-occupation with explanations of meaning in terms of referring, denoting, naming, etc. In the paper quoted at the beginning of this section, Quine examines the notion of referring and argues strongly that even in cases of ostensive definition - where the meaning of a word is explained by pointing to what the word refers to - it is in principle impossible to know precisely what is being pointed at. Thus, if I explain ‘rabbit’ to you by pointing to a rabbit, how can you know that I am pointing to the rabbit as an individual in the external world and not to, say the rabbit’s ears, or to the rabbit as an example of a baby rabbit or a female rabbit. This is what he calls the ‘inscrutability of reference’. From this he argues that reference is ‘indeterminate’, in the sense that one cannot know what is being referred to, and that, a fortiori, translation too is indeterminate. In order to avoid the (absurd) conclusion that ‘there is no difference between the rabbit and each of its parts or stages’ he finishes by proposing that we

begin by picturing us at home in our language, with all its predicates and auxiliary devices. This vocabulary includes ‘rabbit’, ‘rabbit part’, ‘rabbit stage’, ‘formula’, ‘number’, ‘ox’, ‘cattle’; also the two-place predicates of identity and difference, and other logical particles. In these terms we can say in so many words that this is a formula and that a number, this is a rabbit and that a rabbit-part, this and that the same rabbit, and this and that different parts. In just those words. This network of terms and predicates and auxiliary devices is, in relativity jargon, our frame of reference, or co-ordinate system. Relative to it we can and do talk meaningfully and distinctively of rabbits and parts, numbers and formulas.

So once again, we are operating within a language, we are ‘acquiescing in our mother tongue and taking its words at face value’. Both these alternative views, then, leave us with some uncomfortable questions about translation. If meaning is language internal, and if we operate meaningfully only within a frame of reference given by a particular language, what happens when we translate?

Quine hints at one possible way out when he talks of ‘predicates of identity and difference, and other logical particles’. Perhaps there is a way out via logic and truth functions: through provision of a formal calculus within which we could model the external world, and such a model would be independent of any particular language. This is the line taken by, for example, Cresswell, in the excerpt below; where he tries to re-formulate the ‘independent entities’ view in terms of possible world semantics:

Let us say that a and b (in different languages) are correct translations, each of the other, if they have the same meaning. For this definition to be viable we require language-independent entities to be the meanings of a and b, so that we can say that the entity which is a’s meaning in its language is the same entity as that which is b’s meaning in its language.

Our theory of semantic competence enables a speaker to match up a and b with sets of possible worlds. Since possible worlds are language-independent he has the ability to tell whether a and b are logically equivalent, and it will be a necessary, though not in general a sufficient, condition

for a sentence a and a sentence b to be (correctly) intertranslatable that they be true in precisely the same set of possible worlds...

What seems likely is that an adequate treatment of these problems in possible-worlds semantics will require the use of theoretical entities which can represent distinct, though logically equivalent, propositions. If these entities are language-independent (as sets of possible worlds are language-independent) then, by treating them as the meanings of sentences in any language, we can say that a and b, in different languages, are correct translations of each other if they have the same meaning.

Since an adequate truth-conditional semantics will have to account for expressions like ‘means the same as’ it follows that, in order to solve its own internal problems, a theory of meaning based on the truth-conditional view of semantic competence offered in this paper will be sufficiently constrained to test the correctness of any synonymy claim in a single language; and granted the use of language-independent meanings, will therefore be able to test the correctness of any translation between one language and another.

Although this account avoids some (although not all) of the difficulties in the ‘independent entities’ view by offering a definition of them in terms of equivalence with possible worlds, it critically relates both meaning and translation to the formulation of truth conditions. This seems to me to be a mistaken view of what translation is, in that it neglects its essential linguistic character. Presumably, ‘there is a chair in this room’ and ‘there is an article of furniture in this room known as a chair’ have the same truth conditions, but it seems implausible to regard either one as a translation of the other. Even if this difficulty could be overcome, we are, in a rather similar way, left with the problem of formulating truth conditions for utterances containing semantically similar but translationally different terms in such a way that they can be distinguished. (For example, ‘kill’, ‘slay’, ‘assassinate’, ‘murder’.) I find it difficult to imagine a formal apparatus which would allow us to do this in any very satisfactory way.

Thus, it seems that we are left with the notion that meaning is only describable in terms of a word’s use within a language; there is no escape from words towards language independent entities. Yet we know that people do in fact translate, and that sometimes, at least, translation seems to be satisfactory. I want to suggest that the trick to understanding how this can be so is to take seriously the idea that a particular language imposes a structure on the world, rather than reflects a structure which is independently given; then, since the world as perceived through one language has much in common with the world as perceived through a different language, the structures imposed will sometimes, although not always, correspond in quite important ways. Learning a language does not involve taking one language and learning to express its terms and relations in terms of a second language, but involves learning a new world. Once this is done, it then becomes possible to identify correspondences between the two worlds created, as it were, by the two languages. On this thesis, learning a second language is not very different from learning a first language. Some support can be found for this in the empirical fact that adults learn languages with much less facility than do children: if language learning were simply a matter of learning a different,

but equivalent, vocabulary in which to express relations with which one was already familiar, then increased experience with one language gained through age should make it easier to learn the second language. Similarly, there is some (anecdotal rather than systematic) evidence which, in going against the language independent meanings thesis, tends to support the view put forward here: when bi- or multi-linguals attempt to recount a story or joke heard in one language in another, they very frequently hesitate to search for translations, suggesting that they remember the story in the language in which they heard it. (There is, however, some psychological evidence that humans do not seem to store surface language; see, for example, Johnson-Laird, 1974.) I do not know how much the question of mono-linguality vs. multi-linguality affects the issue or even whether the contradiction is, in fact, very deep).

A translator, then, identifies correspondences between the structures imposed by two languages; where such a correspondence exists, translation is possible, where it does not, only an approximation can be made. On reflection, this should not seem very shocking. There are very well-known cases where languages do not correspond; cases of lexical holes, for example, where one language simply does not have a vocabulary item corresponding to a vocabulary item in another language, or cases like colour words where there is strong evidence that different languages carve up the world in different ways. It does however have some consequences for the status of a variety of tools used within machine translation systems.

### 3 Tools for Semantic Analysis and the Search for Universals

In this section I want to look at a number of tools used within machine translation systems and ask what is their conceptual status in the light of the foregoing.

The clearest case is perhaps that of frames or scripts, when these are used as a way to drive the analysis rather than (or as well as) as a representation to be aimed at as the output of analysis. The classic case is to take some stereotypical situation or event, describe it in a formal calculus of some sort, and then carry out the analysis by seeking to relate elements of the input text to elements in the stereotypic description. The result will be a representation of the text, usually in terms of the stereotypic situation or event. In a machine translation system, the output translation is then based on this representation. (Metzing, 1979 is a collection of papers on the use of frames where the interested reader will find much more detailed description.) Despite some rather rash claims to the contrary (not the responsibility of the originator of the proposal), such organisations of knowledge about situations or events are very clearly bound to a particular culture and therefore to a particular language. Thus, translation can only succeed when there is a correspondence, in the sense of the preceding section, in the perceptions of the stereotyped event. Otherwise, the correspondence may be close enough to allow a frame for one language to be mapped onto a corresponding frame for the other. This would be the case, for example, with Minsky's birthday party frame if the system were translating between American English and French. (The typical party games are different, the food is different and so on, but the mapping could be done). In the extreme case, no correspondence will exist, and translation, as such, will be impossible. It is not too difficult, for example, to imagine a society in which children simply do not have birthday parties. Here, if a description of a birthday party occurred in a text to be translated, it

might be possible to add some explanation along the lines of ‘In America, once a year on the date on which they were born, children are invited to a feast by their parents. This is called a birthday party.’ But doing this is very clearly adding to the target language and its concepts, rather than taking it as given and simply moving from one language to another.

Equally clearly, the frames identified as aids for the analysis of a language can make no serious claims to universality in principle. It would perhaps be possible to imagine that some set of frames contingently applied to all languages (indeed, it was once, not very seriously, suggested that birth-life-death might constitute such a frame), but this would only be by accident rather than by definition within the theory.

Frame systems are rather similar in spirit to systems based on model theoretic semantics. Within such systems, typically, the independent existence of some state of affairs is assumed as the entity denoted by a semantic representation, and the semantic representation is isomorphic to the state of affairs. (Most logical grammars provide examples of this type of semantics.) However, the state of affairs modelled is often some feature of the world, rather than events or situations. A good example can be found in much of the recent work in artificial intelligence and in some machine translation systems on the modelling of time. An attempt is made to set up a model expressible in terms of a formal calculus, relatively independent of a specific language (I say relatively because most model-theoretic representations contain language specific words, like ‘before’ or ‘after’; however they may contain parts such as their structure and parts of the vocabulary which are intended to be interlingual, and, more importantly, they receive an interpretation in terms of the semantic domain they model.) Analysis of the input text then involves extracting from it those elements relevant to the semantic domain in question and mapping the information thus retrieved onto the formal calculus of the model.

Here the question of universality is more difficult. The issue turns not so much on the existence or non-existence of independent meanings as on the existence or non-existence of something like the Kantian universal categories: aspects of the world or of thinking which no language can escape talking about, and where, even if different languages talk about these categories in different ways, it is possible to set up a model comprehending and subsuming their different ways. A strict Wittgensteinian approach would, I think, force one to give this notion up, and there is indeed some evidence that we should do so (Oatley, 1977). It is perhaps a superstitious dread of being set adrift in a formless world that makes us reluctant to do so. Two points however are clear. First, it is extremely unlikely that all of language can be dealt with in this way. To see this, one only has to think of the difficulty of accommodating discourse phenomena or pragmatic factors into a model, or of the intricacy (impossibility?) of dealing with intention. Second, even if a model of some category can be set up, if we admit that different languages may talk about the model in different ways, the existence of the model does not in itself guarantee translatability between the languages. Once again, the only cases in which translatability could be guaranteed would be those cases in which we had a correspondence between the two languages in the sense of the first section.

Quite direct claims for universality have been made by the proponents of conceptual dependency theory: ‘Conceptual Dependency theory is intended to be an interlingual meaning representation. Because it is intended to be language free, it is necessary in our

representation to break down sentences into the elements that make them up. In order to do this it is necessary to establish a syntax of possible conceptual relationships and a set of conceptual categories that these relate' (Schank & Rieger, 1974).

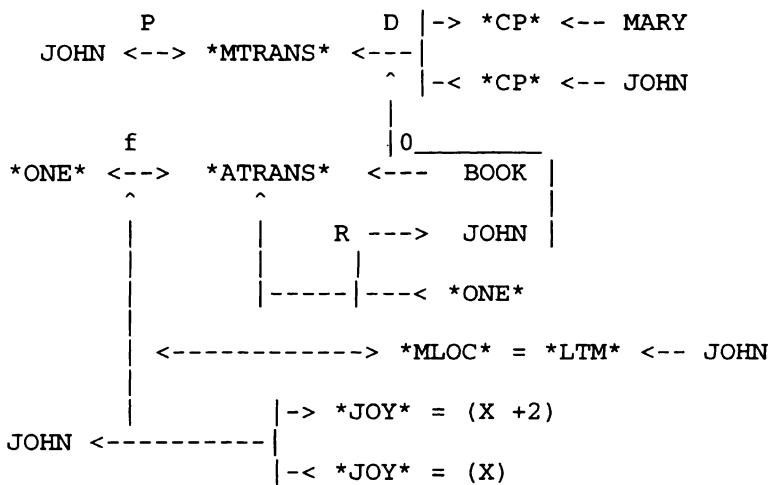
The primary conceptual categories are taken to be ACTs, real world actions, of which a small number are claimed to be primitive. (Twelve in the paper taken here as a basic reference: the number varies slightly in other papers.) Other conceptual categories, for example 'real world objects', 'attribute of actions', 'times', 'locations' are postulated, but are not worked out in any great detail. The theory concentrates on the set of primitive actions. With each action is (obligatorially) associated a sub-set of the conceptual cases (objective, recipient, directive, instrumental, where instrumental is itself a complete propositional structure) and a set of inferences. The inferences occur automatically, but are not guaranteed to be correct. Thus, within conceptual dependency theory, the meaning of 'John told Mary that he wants a book' will not only include John's transmitting to Mary that he would be happier if someone gave him a book, but will also include that John will also be happier if he can transmit information from the book to himself: i.e. the inference is made that John wants a book because he wants to read it.

The representation theory of conceptual dependency and the inferencing techniques associated with it can be considered independently, as can be seen from Rieger's thesis, which is mainly concerned with distinguishing different types of inference (Rieger, 1974). Here we shall consider only the status of the primitive acts and of the 'conceptualizations' (i.e. propositional structures) of which they form a part.

First, and most obviously, primitive acts themselves cannot be considered to be independent meanings of words: the mere fact that (at least) all the verbs of English can, it is claimed, be mapped onto twelve or so primitive acts makes this obvious.

Nor would Schank argue differently. In Schank (1975) he postulates the existence of two kinds of memory. (Conceptual dependency theory was originally developed as a model of human memory: its use inside natural language understanding systems was seen as a way of testing the theory, rather than as an independent enterprise. Later workers, however, have sometimes taken over the theory as a theory about language processing.) One of these is conceptual memory, which is structured in the conceptualizations described briefly above. The other is a lexical memory, which is said to contain 'all of the information about words, idioms, common expressions etc.' and which 'links these to nodes in a conceptual memory, which is language free'. Furthermore, the lexical memory seems to have some structure, since, in discussing the notion of 'superset' in memory (in relation to semantic network representations such as those of Collins & Quillian, 1969, where 'bird', for example, is represented as a superset of 'canary', 'ostrich' and so on), Schank claims that the number of such supersets is very small and that they are 'mostly artificial constructs with definitions in lexical memory'. It is not clear, though, where or how this lexical memory comes into being: Schank (1974) seems to contrast it with conceptual memory; 'Once we change semantic memory by separating out lexical memory, we are left with a set of associations and other relations between concepts that could only have been acquired by personal experience'. This would seem to imply that lexical memory is not acquired by personal experience, but nothing more is said, and it would perhaps be rash to push too far on the basis of a single sentence.

On the other hand, the conceptualizations themselves claimed to represent the meaning of sentences, and thus could be expected to explicate or at least subsume the meaning



(i.e., the conceptualization to the left has mental location John's LTM)

of the individual words. That they do not in fact do so in any normal sense of 'meaning' can be seen by considering the conceptualization for 'John told Mary that he wants a book' (from Schank & Rieger, 1974).

There is no need to go into detail. The essential point is that two primitive acts, ATRANS and MTRANS are involved, the first involving 'the transfer of an abstract relationship such as possession, ownership or control', the second 'the transfer of mental information between animals or within an animal'. It is difficult to see how, with the limited apparatus offered, it would be possible to distinguish between different ways of telling Mary, or of being given a book. If John wrote to Mary, one could perhaps capture this by including an instrument case: but if he hinted it, or informed her that ..., or insisted to her that ..., how could this be represented? Thus, conceptual representations of this sort actually carry less meaning than the sentences they represent. Schank, however, does not base his argument on meaning but on 'information': 'Information is not lost by the use or primitive ACT's' (Schank, 1975).

A defendant of the theory might argue therefore that we were being finicky: that identifying and preserving all important information was all that mattered. The claim then would be that it was possible to set up a formal model which captured all information necessary to communicate what we might call kernel meaning - the bare bones without any of the niceties of nuance or of subtlety. Whether or not such a model were meaning-preserving would then, of course, be very contentious, and would depend critically on a definition of what the essential information was; which, in its turn - and this is logically prior - depends on a belief that it is, in some way, possible to identify, language independently, what has to be communicated.

Schank himself seems to believe this to be possible, although he does not claim that there is some fixed number of ACT's which will constitute the 'right' set:

There is no right number of ACTs. It would be possible to map all of language into combinations of mental and physical MOVE. This would, however, be extremely cumbersome to deal with in a computer system. A larger set (several hundred) would overlap tremendously causing problems in paraphrase recognition and inference organization. The set we have chosen is small enough not to cause these problems without being too small. Other sets of the same order of magnitude might do just as well.

All of this, it seems to me, is to make the explanation of meaning more problematic rather than less. Now it is not even possible to look at how a word is used within a language: instead we have to search for some common element in the use of a group of words (perhaps deciding rather arbitrarily what the members of that group are), define that common element and embed it into an information structure. Having done so, we can reasonably ask what that information structure means. Only two answers seem possible. The first is an answer in terms of some set of operations performed on that structure by, say a computer program, in which case the adequacy of the structure is, presumably, to be judged by the output of the program. To do this, some metric must be set up in terms of which the judgment can be made - and this begins to sound suspiciously like a vicious circle, in that we are back to asking whether the program mimics use in ordinary language. The second is an answer already in terms of ordinary language, without passing through explanation in terms of the computer program, when the circle becomes evident.

Since conceptual dependency was originally developed as a memory model, another justification could be sought in the psychological reality of the conceptualizations and their organization. This too seems extraordinarily difficult to prove, since it involves comparing two unlike objects. The structure of the brain and its workings is still very largely a mystery. The designers of experiments designed to show psychological reality are therefore driven back on an experimental design whereby some particular input is claimed to predict some particular output. Unfortunately, even when the prediction is correct, such an experiment can say little, if anything, about how the output was obtained. An abacus and a computer can be claimed to take similar inputs and produce similar outputs, but no-one claims that they do it the same way.

As far as translatability goes, it should be clear that conceptual dependency theories simply cannot be used to translate. They are, by their nature, reductionist, in that they lose much of what is normally considered meaning: even if we take cases where not much is lost, 'Jean a acheté une voiture' cannot be translated as 'Someone sold John a car'. In other words, at best such theories lead to paraphrase systems. At worst they lead to re-expressing the full vocabulary of French, say, in the eight hundred or so words of basic English.

Associated with the primitive acts of conceptual dependency theory are, as we have already seen, four conceptual cases. The use of deep cases had become quite common in language processing systems since the appearance of Fillmore's seminal paper in 1968. (Although it seems that Japanese linguists - perhaps not surprisingly given the structure of Japanese - were working along similar lines well before Fillmore). Fillmore makes

a distinction between surface case, marked in some languages by inflection (e.g. nominative, accusative, genitive, dative, ablative in Latin), by prepositions in others (e.g. English), by post-position particles in others (e.g. Japanese), and ‘deep’ case, which is held to capture the semantic role of the participant in a predicate. In terms of deep case ‘John’, ‘the door’ and ‘the key’ preserve their semantic role in all of the following sentences, despite the differences in surface structure:

- ‘John opened the door with the key’
- ‘The key opened the door’
- ‘The door opened with a key’
- ‘The door opened’
- ‘The door was opened by John’
- ‘The door was opened by John with a key’ etc.

Both for Fillmore and for Schank, the number of deep cases is fairly small. Workers in machine translation have tended, on the other hand, to work with much larger case sets (typically of around twenty to thirty cases: the increase in the number of cases comes largely from finer classification of circumstantial to take in roles like ‘result’, ‘cause’, ‘concessive’ in addition to the time and location cases).

Fillmore himself claimed only that the notion of deep case was universal, in the sense that, in Chomskyan terms (Chomsky, 1965) a notion of a case formed part of the base structure: ‘...what is needed in a conception of base structure in which case relationships are primitive terms of the theory and in which such concepts as “subject” and “direct object” are missing’. However, he makes no claim for the universality of particular case sets: ‘My claim is, then, that a designated set of case categories is provided for every language, with more or less specific syntactic, lexical, and semantic consequences, and that the attempt to restrict the notion of “case” to the surface structure must fail’.

(Fillmore did, though, rather hope that on investigation, some universality might emerge: ‘It seems to me that if there are recognizable intrasentence relationships of the types discussed in studies of case system... that if these same relationships can be shown to be comparable across languages and that if there is some predictive or explanatory use to which assumptions concerning the universality of these relations can be put, then surely there can be no meaningful objection to using the word case, in a clearly understood deep-structure sense, to identify these relationships’. We shall return to the question of empirical investigation in the next section).

Some support for the idea that different languages rely on different case sets can be found also in the empirical fact that the Japanese national machine translation project, which relies heavily on the use of deep case, uses a case set for the analysis of Japanese which differs slightly from that used in the same project for the analysis of English (Nagao et al. 1985). Even where the deep case intuitively seems almost to correspond to one of the universal categories (as we called them earlier), there is evidence that different languages give a somewhat different semantic content to the case roles. Consider, for example, space relations in English and French, where the choice of preposition is determined by quite different factors:

	in (cities, countries, enclosed spaces)
	on (islands, mountains, streets)
	at (buildings)
vs	à Paris                    en Avignon (phonetic)
	en France                au Japon (gender)
	en Corse                  sur l'île d'Elbe (political status)

(This example is due to H. Somers).

In view of what was said in the first section, it should not greatly surprise us to find doubts about the a priori universal status of deep cases or semantic relations. They could only be universal if all languages structured the world in the same way, which seems unlikely if language is taken as imposing structure on the world rather than as reflecting a world already structured. As with other aspects of meaning, sometimes there will be a strict correspondence in the relationships perceived, sometimes not. Translation will consist, as before, in identifying and using the strict correspondences, learning to massage the near correspondences and finding, where possible, other ways of expressing the mis-matches.

One final tool of semantic analysis should be considered before we leave this topic: the use of semantic markers. Semantic markers can be thought of in two quite distinct ways, although there is a frequent and unfortunate confusion between the two. Both can be found in the original Katz and Fodor proposal. First, semantic markers are intended to represent the meaning of individual words. Second - and, conceptually, this is quite a distinct notion - semantic markers are used to encode selectional restrictions, and thus to inhibit the co-occurrence of certain words. As an example, take the dictionary entry for "ball"

Ball → concrete noun → (social activity) → (large) → (assembly) → [dance]  
 Ball → concrete noun → (physical object) → [sphere]  
 Ball → concrete noun → (physical object) → [cannon-ball]

(From Katz and Fodor, 1963.)

The objects in round brackets are the semantic markers, the objects in square brackets 'distinguishers'. The basic notion of selectional restrictions is very simple: it rests on the observation that some word senses which are fundamentally predicates or relations may only predicate on or relate objects having certain properties. Thus, the verb 'kick' for example, in the ordinary sense of propel by use of a foot or a hoof requires an (animal) as subject and a (physical object) as object. So, in the sentence 'The page kicks the ball', the sense of page as the page of a book is ruled out by the requirement that the subject must have a marker (animal), and the first of the three senses of ball in the example is ruled out by the requirement that the object must have (physical object).

The distinguishers are intended to contain the remnant of the meaning of a particular word sense which is not already accounted for by the semantic markers in the entry for that sense. They do not, though, interact with the selectional restrictions, which operate only in terms of the markers. However, markers and distinguishers together should capture the meaning of the word.

If we now neglect the distinction between markers and distinguishers, and consider only the two mechanisms, that of defining the meaning of the word, and that of serving as clues for disambiguation and thus inhibiting certain co-occurrences, we see immediately that the two functions are quite different.

Taken as a means of defining the meaning of a word, markers become semantic primitives, combinations of which produce particular meanings. This is a variation on the ‘independent entities’ theory of meaning, where some set of primitives is taken as primary (and independent) and ‘meanings’ constructed out of them. As such, it is open to all the objections levelled against that theory.

Wilks (1975a & 1975b) has put forward a version of markers-as-meanings which avoids this objection. There he argues that semantic primitives, in such a scheme, have exactly the same status as words of the language, and that, in fact, they coincide with words of the language in everything except that they are words accorded an extra privilege in that verbal explanation of the meaning of other words is only allowed in terms of the privileged words.

If we take this position, it follows that there can be no ‘right’ set of semantic primitives. To quote Wilks (1975b): ‘It follows from this that there can be a variety of primitive languages for semantic descriptions, no one necessarily better or worse than any other, any more than my vocabulary is better or worse than yours if I know 100 English words you don’t, and you know 101 that I don’t. In the case of each primitive vocabulary, the only ultimate test will be the success or failure of linguistic computations that make use of it’. He goes on to remark that there are some limits to the parallel between a primitive vocabulary and ordinary language, for example that a primitive vocabulary should not have synonymous primitives, and earlier in the paper suggests other ‘operational’ criteria for the selection of a primitive vocabulary.

There are further objections to the notion of defining a single, universal, correct set of primitives. The first is already implicit in the quotation from Wilks above. If a set of primitives is to be universal, each primitive must be interpreted in the same way. But how is this to be achieved? It seems difficult to offer a definition, since primitives are primitive by definition, and are those terms in a theory which are no longer defined. Therefore, all that could be offered would be an explanation of the use of the primitive, and if this is offered in words or by examples first no consistency of interpretation can be guaranteed and - worse - we are again in the vicious circle where primitives are defined by words and words by primitives. Secondly, if we accept that there can be no objective correlative of a would-be semantic primitive, how could anyone using a set of primitives ever know that he was using it in the correct way? In the end, the interpretation of a set of primitives can only be given by the use to which they are put within some formal system. This point too we shall return to in the next section.

The arguments offered here against the possibility of defining a correct set of primitives hold equally strongly for the definition of a set of markers to be used as clues in disambiguation if it is assumed that for every language, or even within one language, there will be a single correct set. This is sometimes somewhat obscured by the fact that the markers most commonly found in systems using them in this way tend to constitute a rather small set using words like ‘animate’, ‘physical object’ which reflect common strong correspondences between languages. As soon as the set is enlarged, the temptation to believe in a single universal set recedes.

One final remark: if we accept Wilks' thesis that primitives are no more than ordinary words accorded a privileged status, then of course they cannot be universal across languages. 'Animate' is an English word, and enters into the nexus of English language use; 'animé' is a French word entering into the nexus of French language use. The two uses do not correspond.

## 4 The Proper Place of Semantics in Machine Translation

The reader might be forgiven for believing that I am advocating the abandonment of semantic tools for use in machine translation systems. In fact, this is not so, I am advocating only that they are seen in a correct perspective, and that system designers do not prejudice their own intellectual enterprise by trying to insist on a universality which does not exist. In this section, I want to put forward some suggestions about the appropriate use of semantic tools. Apart from some very brief remarks on the other techniques discussed in the last section, I shall concentrate mainly on the use of case and of semantic features.

Let me start by insisting once again on the notion of translation as a linguistic undertaking, where the primary task is to find words of one language, embedded in an appropriate syntactic structure, which capture as closely as possible the correspondence between the world of the target language and the world of the source language, this latter itself described via the mechanisms of the source language, with all its richness of vocabulary and its subtlety of structure.

Providing that this is not forgotten, with the result that we finish up with an impoverished paraphrase rather than a translation, there is an obvious place for the use of frames whenever there is a very strong correspondence between the target language world and the world of the source language. There are, of course, other technical problems in the use of frame-driven systems which have not been discussed in this paper which would then become relevant, but they do not matter for the point being made here.

The case of conceptual dependency type theories is rather more difficult. In the preceding section, I have suggested that they are based on a reductionist hypothesis of meaning, in that whole classes of verbs are mapped onto the primitive acts, with a consequent automatic loss of meaning. This would seem to make them unsuitable for use in a translation system if the conceptual dependency representation is taken as an interlingua (of sorts) through which translations must pass. (Recent work using conceptual dependency within the Yale group has tended to concentrate on frame-like systems, using conceptual dependency representations as the smallest building blocks of the frames. This may make it seem that this paragraph is inconsistent with the last. But, there is, of course, no essential intimate connection between frame systems and conceptual dependency representations, as witness the fact that most frame systems use quite different, and quite varied, representations.)

On the other hand, although we have not discussed the issue much in this paper, conceptual dependency representations are often used as the basis for inference making. Perhaps, then, they could be useful in those cases where inference becomes important, for example to resolve difficult problems of pronoun reference. But I would want to claim that their use should be restricted to the solution of such problems, and that they

should not be seen as the primary tool around which a system should be built, thus replacing the nuts and bolts of standard linguistic analysis.

Model theoretic semantics too may find a place for sub-parts of a system where it is possible to set up a formal calculus onto which some aspects of the language(s) treated by the system can be mapped, but cannot serve as the basis of a system in general.

The difference between what I am suggesting here and the proposals normally found in the literature is this: normally, the use of frames, of conceptual dependency like theories or of model theoretic semantics is seen as an alternative to a conventional linguistic analysis based on syntax. Here I am suggesting that such tools should work in co-operation with a more conventional model. By this, I do not want to suggest that the only possible semantics is an interpretative semantics, whereby semantic representations are considered to be straightforward projections of syntactic representations. Typically, within such theories, the semantic representation is derived from autonomously defined syntactic structures, there are quite severe constraints on the rules which define the mapping from one to the other, and in consequence, the semantic representations tend to be quite similar to the syntactic representations. (For examples of this type of theory, see Chomsky, 1981, Jackendoff, 1983, Bresnan, 1982.) Since, within a machine translation system, I take it to be the main task of analysis to neutralize the differences between languages as much as possible, and since, within such theories, the semantic interpretation tends to reflect the syntactic analysis, which, in its turn, reflects differences which could, otherwise, be neutralized, I take interpretive semantics to be of only limited interest within a machine translation system. Rather, I am suggesting some kind of 'blackboard' approach, where different kinds of analysis contribute their conclusions to a common pool, and the final result is constructed on the basis of what is found in the common pool.

The two remaining semantic tools of those discussed in the last section have normally been perceived as additions to, or aids towards, a conventional linguistic analysis, rather than as a replacement for it. Most commonly, deep cases are seen as a way to neutralize the contingent syntactic variations of surface structure: a constituent analysis of some sort is performed, and the constituents thus identified mapped into a set of case roles, which are intended to be common to the two or more languages being treated within the particular machine translation system. This indeed seems to be the most appropriate use for a set of deep cases. The difficulty, of course, comes with the definition of an appropriate set, especially if we accept the argument of the last section that no correct universal set can be defined *a priori*. To regard this as an insurmountable obstacle seems to me, however, to be unnecessarily defeatist, since it is perfectly possible to define an operational procedure whereby correspondences between cases can be found across languages. Such a procedure would consist in establishing a set of cases for each language separately, taking as criteria for an appropriate set such factors as results of previous work on that language, the possibility of ensuring consistent assignment (by men and by machines), the distinctions made within that language, and so on. Transfer would then involve quite explicitly mapping between the deep case roles of one language and those of another. If it were found, on the basis of such an empirical investigation, that the mapping between any two case roles was always one to one (or even, with a very slight adjustment of the case set, could be made so), it could then be taken that a relationship common to both languages had been found, and explicit transfer could be

dropped.

(As an historical, and perhaps provocative note, I believe Fillmore himself to have been assuming the necessity of just such an empirical investigation, as shown by the quotation in the last section and by sentences such as 'I am going to suggest below that there are many semantically relevant syntactic relationships involving nouns and the structures that contain them, that these relationships ... are in large part covert but are nevertheless empirically discoverable, that they form a specific finite set, and that observations made about them will turn out to have considerable cross-linguistic validity. I shall refer to these as "case relationships"'). One of the beauties of a transfer-based machine translation system is that it provides a clear framework in which such an empirical study can be carried out.

It is not nearly so clear that any similar empirical study can be proposed in the case of semantic features. It will, perhaps help to clarify the issue somewhat if we distinguish two possible places in which semantic features might be used. The first would be their use inside some area of language susceptible to modelling in the model theoretic sense described earlier (if such an area exists). If we take as an hypothesis that it is possible to set up such a model, for time say, then it is easy to imagine that part of the model itself would consist in a set of features. In such a case, there is no question of an appropriate or inappropriate set: the model itself determines the features which are part of it. But there is another common use of semantic features which consists in using the features as an aid towards some further result, in carrying out the mapping between the text and the model, say, or as an aid to lexical or structural disambiguation. This case, I would argue, is quite different. Here, the semantic features can be seen most fruitfully as an attempt at semantic categorisation, and thus closely parallel to syntactic or morphological category names. If this is so, then just as syntactic categories receive their interpretation within the linguistic theory which makes use of them - in machine translation terms via the grammar rules in which they are used - so do semantic categories. And just as we expect that different theories will make use of different syntactic categories, and do not ask for there to be one universal set, so we should expect that different theories will make use of different semantic categories, and not ask for or expect universality here either. We would then take Wilks' proposal, whereby semantic primitives are regarded as a primitive vocabulary, as a proposal for establishing a preliminary set of semantic features within a particular theory as well as within a particular language, and would add to his criteria of lack of redundancy and disjointness the further criterion of usefulness within the theory. The semantic categories would then serve as aids to establishing a correct analysis and a correct generation of a specific language, and would serve cross-linguistically only as extra information towards making correct choices in transfer.

## 5 Conclusion

This paper, starting from philosophical considerations, has discussed the status of a variety of tools used in natural language understanding systems in general and in machine translation systems in particular, with the aim of determining their most appropriate use within such systems. Relatively little has been said that will be new to philosophers or linguists: the point in saying it has been to emphasize to workers in machine translation and in computational linguistics in general the importance of being clear about the nature

of the intellectual tools they use and about their reasons for using them. That point will remain valid even if every argument or conclusion contained in the present paper proves to be mistaken.

### Acknowledgments

Although the views expressed in this paper are my own, I would like to thank my Eurotra colleagues, perhaps especially those who would disagree with them, for the discussions and arguments on this topic which have pushed me into trying to gather many arguments into one place. In particular, I would like to thank Kirsten Falkedal, Harry Somers, Frank Van Eynde and Cornelia Zelinsky.

## References

- [1] Austin, J.L., 1961. "The Meaning of a Word," *Philosophical Papers*. Oxford University Press.
- [2] Austin, J.L., 1962. *How to do Things with Words*. Harvard University Press, Cambridge, MA.
- [3] Bresnan, J. (ed.), 1982. *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA.
- [4] Caton, C.E., 1971. "Philosophy: Overview," in Steinberg & Jakobovits, 1971.
- [5] Chomsky, N., 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.
- [6] Chomsky, N., 1981. *Lectures on Government and Binding*. Foris Publications, Dordrecht.
- [7] Colling, A.M. and Quillian, M.R., 1969. "Retrieval Time for Semantic Memory," in *Journal of Verbal Learning and Verbal Behavior*, 8.
- [8] Cresswell, M.J., 1978. "Semantic Competence," in Guenthner and Guenthner-Reuter, 1978.
- [9] Fillmore, C.J., 1968. "The Case for Case," in Bach and Harms (eds.), *Universals in Linguistic Theory*, Holt, Rinehart and Winston, New York.
- [10] Guenthner, F. and Guenthner-Reuter (eds.), 1978. *Meaning and Translation*. Duckworth.
- [11] Jackendoff, R., 1983. *Semantics and Cognition*. MIT Press, Cambridge, MA.
- [12] Johnson-Laird, P., 1974. "Memory for Words," *Nature*.
- [13] Katz, J. and Fodor, J.A., 1963. "The Structure of a Semantic Theory," *Language*, 39.

- [14] Metzing, D. (ed.), 1979. *Frame Conceptions and Text Understanding*. De Gruyter, Berlin.
- [15] Minsky, M., 1981. “A Framework for Representing Knowledge,” in Haugeland (ed.), *Mind Design*. MIT Press, Cambridge, MA.
- [16] Nagao, M., Tsujii, J.-I. and Nakamura, J.-I., 1985. “The Japanese Government Project for Machine Translation,” *Computational Linguistics*, 11(2-3).
- [17] Oatley, K.G., 1977. “Inference, Navigation and Cognitive Maps,” in Johnson-Laird and Wason (eds.), *Thinking*. Cambridge University Press.
- [18] Quine, W.V., 1960. *Word and Object*. MIT Press, Cambridge, MA.
- [19] Quine, W.V., 1971. “The Inscrutability of Reference,” in Steinberg and Jakobovits, 1971.
- [20] Rieger, C., 1974. *Conceptual Memory*. Unpublished Ph.D. Thesis, Stanford University.
- [21] Ryle, G., 1953. “Ordinary Language,” *Philosophical Review*, LXII.
- [22] Ryle, G., 1957. “The Theory of Meaning,” in C.A. Mace (ed.), *British Philosophy in the Mid-Century*.
- [23] Schank, R.C., 1974. *Is There a Semantic Memory*. ISSCO Working Paper No. 3.
- [24] Schank, R.C., 1975. “The Theory of Meaning,” in C.A. Mace (ed.), *British Philosophy in the Mid-Century*.
- [25] Schank, R.C. and Rieger, C.J., 1974. “Inference and the Computer Understanding of Natural Language,” *Artificial Intelligence* 5(4).
- [26] Somers, H.L. (Forthcoming). *Valency and Case in Computational Linguistics*. Edinburgh University Press.
- [27] Steinberg, D.D. and Jakobovits, L.A., (eds.) 1971. *Semantics*. Cambridge University Press.
- [28] Wilks, Y., 1975a. “Primitives and Words,” in Schank and Nash-Webber (eds.) *Proceedings of the Workshop on Theoretical Issues in Natural Language Processing*. Bolt, Beranek & Newman, Cambridge, MA.
- [29] Wilks, Y., 1975b. *Seven Theses on Artificial Intelligence and Natural Language*: ISSCO Working Paper No. 17.
- [30] Wittgenstein, L., 1953. *Philosophical Investigations*. Blackwell, London.
- [31] Wittgenstein, L., 1958. *The Blue and Brown Books*. Blackwell, London.

# Developing a Natural Language Interface to Complex Data\*

Gary G. Hendrix, Earl D. Sacerdoti,  
Daniel Sagalowicz and Jonathan Slocum  
*e-mail: slocum@cs.utexas.edu*

## Abstract

Aspects of an intelligent interface that provides natural language access to a large body of data distributed over a computer network are described. The overall system architecture is presented, showing how a user is buffered from the actual database management systems (DBMSs) by three layers of insulating components. These layers operate in series to convert natural language queries into calls to DBMSs at remote sites. Attention is then focused on the first of the insulating components, the natural language system. A pragmatic approach to language access that has proved useful for building interfaces to databases is described and illustrated by examples. Special language features that increase system usability, such as spelling correction, processing of incomplete inputs, and run-time system personalization, are also discussed. The language system is contrasted with other work in applied natural language processing, and the system's limitations are analyzed.

## 1 Introduction

In dealing with a very large database (VLDB), which is perhaps distributed among multiple computers with different database management systems (DBMSs) on remote sites, a central problem faced by would-be users is that of formulating queries in terms communicable to the system.

It is usually the case that business executives, government officials, and other decision makers have a good idea of the kind of information residing in their databases. Yet to obtain the answer to a particular question, they generally need to employ the services of a technician who works with the database on a regular basis and who is thoroughly familiar with its file structure, the DBMSs on which it resides, how it is distributed among various computer systems, the coded field names for the data items, the kinds of values that different fields are expected to contain, and other idiosyncrasies.

The technician must understand the decision maker's question, reformulate it in terms of the data that is actually stored, plan a sequence of requests for particular items from particular files on particular computers, open connections with remote sites, build programs to query the remote systems using the primitives of the DBMSs of the remote

---

\*This article was originally published in *ACM Transactions on Database Systems* 3(2), June 1978, pages 105-147. At that time, the authors were at SRI International.

systems, monitor the execution of those programs, recover from errors, and correlate the results. This is a demanding, time-consuming, and exacting task requiring much attention to detail. Escalated levels of sophistication are needed as the VLDB increases in size and complexity and as it is distributed over a wider range of host computers.

With the goal of making large, distributed databases directly available to decision makers (while freeing technicians from increasingly tedious details), a group of researchers at SRI International has developed a prototype system that, for many classes of questions, automates the procedures usually performed by technicians. This paper presents an overview of this system, called LADDER (for language access to distributed data with error recovery) (Sacerdoti 1977)<sup>1</sup>, and then concentrates on the particular problem of translating user queries from English into the terms of the database. The other aspects of the LADDER system are presented in greater detail elsewhere in the literature (Morris 1977, Paxton 1977, Sagalowicz 1977). The system was developed as a management aid to Navy decision makers, so examples throughout the paper will be drawn from the domain of Navy command and control.

## 2 System Architecture

The running demonstration system consists of three major components that provide levels of buffering of the user from the underlying DBMSs. The LADDER user can think he is retrieving information from a “general information base” rather than retrieving specific items of data from a set of highly formatted, traditional databases that are scattered across a computer network. The user provides a question about the information base in English; LADDER applies all the necessary information concerning the vocabulary and syntax of the question, the names of specific fields, how they are formatted, how they are structured into files, and even where the files are physically located, to provide an answer.

LADDER’s first component, called INLAND (for informal natural language access to navy data), accepts questions in a restricted subset of natural language and produces a query or sequence of queries to the VLDB as a whole. The queries to the VLDB, as produced by INLAND, refer to specific fields, but make no commitment about how the information in the database is broken down into files.

For example, INLAND translates the question “What is the length of the Kennedy?” into the query

((NAM EQ JOHN#F.KENNEDY)(? LENGTH)),

where LENGTH is the name of the length field, NAM the name of the ship name field, and JOHN#F.KENNEDY the value of the NAM field for the record concerned with the Kennedy.

Queries from INLAND are directed to the second component of LADDER, called IDA (for intelligent data access)(Sagalowicz 1977). In general, a query to IDA is a command list of constraints (such as (NAM EQ JOHN#F.KENNEDY) or (\* MAX LENGTH)) and requests for values of fields (such as (? LENGTH)). INLAND operates by building a (possibly null) fragment of a query to IDA for each lower-level syntactic

---

<sup>1</sup>A glossary of system names precedes the Appendix.

unit in the English input. These fragments are combined as higher-level syntactic units are recognized. At the sentence level, the combined fragments are sent as a command string to IDA.

Employing a model of the structure of the VLDB, IDA breaks down a query against the entire VLDB into a sequence of queries against individual files. Linkages among the records retrieved are preserved so that appropriate answers to the overall query may be composed and returned.

For example, suppose that the database consists of a single file whose records contain the fields

(NAM CLASS LENGTH).

Then, to answer the database query issued above, IDA can simply create one file retrieval program that says, in essence, "For the ship record with NAM equal JOHN#F.KENNEDY, return the value of the LENGTH field." Suppose, however, that the database is structured in two files, as follows:

SHIP: (NAM CLASS...)

CLASS: (CLASSNAME LENGTH...).

In this case the single query about the Kennedy's length must be broken into two file queries. These would say, first, "Obtain the value of the CLASS field for the SHIP record with NAM equal JOHN#F.KENNEDY." Then, "Find the corresponding CLASS record and return the value of the LENGTH field from that record."<sup>2</sup> Finally, IDA would compose an answer that is relevant to the user's original query (i.e., it will return NAM and LENGTH data, suppressing the CLASS-to-CLASSNAME link).

In addition to planning the correct sequence of file queries, IDA must actually compose those queries in the language of the remote DBMSs. Currently the system accesses, on a number of different machines, a DBMS called the Datacomputer (Computer Corporation of America 1975, Farrell 1976), whose input language is called DATALANGUAGE. IDA creates the relevant DATALANGUAGE query by inserting field and file names into prestored templates. However, since the database is distributed over several machines, the DATALANGUAGE that IDA produces does not refer to specific files in specific directories on specific machines. It refers instead to *generic files*, files containing a specific kind of record. For example, the queries discussed above might refer to the SHIP file rather than file SHIP.ACTIVE in directory NAVY on machine CCA-2.

It is the function of the third major component of LADDER to find the location of the generic files and manage the access to them. To carry out this function, the third component, called FAM (for file access manager) (Morris 1977), relies on a locally stored model showing where files are located throughout the distributed database. When it receives a query expressed in generic DATALANGUAGE, it searches its model for the primary location of the file (or files) to which it refers. It then establishes connections over the ARPANET to the appropriate computers, logs in, opens the files, and transmits the DATALANGUAGE query, as amended to refer to the specific files that are being

---

<sup>2</sup>If it is possible to perform multiple file accesses with a single multifile query, IDA will do so.

accessed. If at any time the remote computer crashes, the file becomes inaccessible, or the network connection fails, FAM can recover, and if a backup file is mentioned in FAM's model of file locations, it can establish a connection to a backup site and retransmit the query.

The existing system, written in INTERLISP (Teitelman 1975), can process a fairly wide range of questions against a database consisting of some 14 files containing about 100 fields. Processing a typical question takes less than a second of CPU time on a DEC KL-10 computer. An annotated transcript of a sample session with the system is provided in the Appendix.

We emphasize that the three major components of LADDER each address separate portions of the data access problem. Although they have been designed to work in combination, each component is a separate, self-contained module that independently addresses one aspect of data access. For example, the virtual view of the data that IDA supports for its caller would be of value even without a natural language front end. Likewise, the general technology developed for natural language translation may be separated from the data access problem and applied in other domains.

### 3 The Natural Language Component

With the goal of supplying natural language interfaces to a variety of computer software, we have developed a language processing package, called LIFER (for language interface facility with ellipsis and recursion) (Hendrix 1977b), that facilitates the construction and run-time operation of special purpose, applications-oriented, natural language interfaces. INLAND, the linguistic component of our intelligent interface to distributed data, has been constructed within the LIFER framework. Figure 1 gives some indication of the diversity of language accepted by this system. Below we describe the nature of INLAND and illustrate how it was created using LIFER's interactive language definition facilities. Of course, the examples can show only limited aspects of INLAND. We believe the existing INLAND system to be one of the most robust computerized natural language systems ever developed, accepting a wide range of questions about information in the database (as shown in Figure 1) as well as metaquestions about definitions of database fields and the grammar itself.

#### 3.1 Overview of LIFER

Although work in artificial intelligence and computational linguistics has not yet developed a general approach to the problems of understanding English and other natural languages, mechanisms do exist for dealing with major fragments of language pertinent to particular application areas. The idea behind LIFER is to adapt existing computational linguistic technology to practical applications while investigating and extending the human engineering aspects of the technology. The LIFER system supplies basic parsing procedures and an interactive methodology needed by a system developer to create convenient interfaces (such as INLAND) in reasonable amounts of time. Certain user-oriented features, such as spelling correction, processing of incomplete inputs, and the ability of the run-time user to extend the language accepted by the system through the use of paraphrase, are also included in the LIFER package.

What kind of information do you know about  
Is there a doctor on board the Biddle  
Display all the American cruisers in the North Atlantic  
What is the name and location of the carrier nearest to New York  
What is the commanding officer's name  
Who commands the Kennedy  
What is the Kennedy's beam  
When will the Los Angeles reach Norfolk  
Tell me when Taru is scheduled to leave port  
Where is she scheduled to go  
When will Los Angeles arrive in its home port  
When will the Sturgeon arrive on station  
What aircraft units are embarked on the Constellation  
To which task organization is Knox assigned  
Where is the Sellers  
Where is Luanda  
What is the next port of call of the Santa Inez  
When will Tarifa get underway  
Which convoy escorts have inoperative sonar systems  
When will they be repaired  
Which U.S. Navy DDGs have casreps involving radar systems  
What Soviet ship has hull number 855  
To what class does the Soviet ship Minsk belong  
What class does the Whale belong to?  
What is the normal steaming time for the Wainwright from Gibraltar to Norfolk  
What American ships are carrying vanadium ore  
How far is it to Norfolk  
How far away is Norfolk  
How many nautical miles is it to Norfolk  
How many miles is it to Norfolk from here  
How close is the Baton Rouge to Norfolk  
How far is the Adams from the Aspro  
What is the distance from Gibraltar to Norfolk  
What is the nearest oiler  
What is the nearest oiler to the Constellation  
How far is it from Naples to 23-00N, 45-OOW  
What is the distance from the Kitty Hawk to Naples

Figure 1: Sample of acceptable inputs to LADDER  
Figure 1 is continued on the next page

How long would it take the Independence to reach 35-00N, 20-00W  
How long is the Philadelphia  
How long would it take the Aspro to join Kennedy  
What is the nearest ship to Naples with a doctor on board  
What is the nearest USN ship to the Enterprise with an operational air search radar  
What is known about that ship  
How many merchant ships are within 400 miles of the Hepburn  
What are their identities and last reported locations  
What cargo does the Pecos have  
Who is CTG 67.3  
What are the length, width, and draft of the Kitty Hawk  
To whom is the Harry E. Yarnell attached  
What type ships are in the Knox class  
Where are the Charles F. Adams class ships  
What are their current assignments  
What subs in the South Atlantic are within 1000 miles of the Sunfish  
What is the Kittyhawk doing  
How many USN asw capable ships are in the Med  
Where are they  
What are their current assignments and fuel states  
What ships are NOT at combat readiness rating C1  
When will Reeves achieve readiness rating C1  
Why is Hoel at readiness rating C2  
When will the sonar be repaired on the Sterett  
What ships are carrying cargo for the United States  
Where are they going  
What are they carrying  
When will they arrive  
Where is Gridley bound  
Which cruisers have less than 50 per cent fuel on board  
Where are all the merchant ships  
When will the Kitty Hawk's radar be up?  
What ships are in the Los Angeles class  
What command does Adm. William have  
Under whose opcon is the Dale  
Show me where the Kennedy is!  
What ship has hull number 148?  
What is the next port of call for the South Carolina?

Fig. 1 (continued from previous page)

Are doctors embarked in the Kawishiwi  
What kind of cargo does the Francis McGraw have?  
What air group is embarked in the Constellation?  
What do you know about the employment schedule of the Lang?  
Which systems are down on the Kitty Hawk  
What ships in the Med have doctors embarked?  
How many ships carrying oil are within 340 miles of Mayport?  
What sub contacts are within 300 miles of the Enterprise?  
List the current position and heading of the US Navy ships in the Mediterranean every 4 hours  
What is the status of the Enterprise's air search radar?  
Where is convoy NL53 going  
What convoy is the Transgermania in  
How many embarked units are in Constellation  
What ships are in British ports  
What U.S. ships are within 500 miles of Wilmington?  
What U.S. ships faster than the Gridley are in Norfolk  
What is the fastest ship in the Mediterranean Sea  
How close is that ship to Naples?  
What is its home port  
Print the American cruisers' current positions and states of readiness!  
How is the Los Angeles powered  
What ship having a normal cruising speed greater than 30 knots is the largest  
Display the last reported position of all ships that are in the North Atlantic  
When did the Endeavour depart the port of New York  
What nationality is the ship with international radio call sign UAID  
What ports are in the database  
What merchant ships are enroute to New York and within 500 miles of the Saratoga  
To what country does the fastest sub belong?

Fig. 1 (continued from previous page)

LIFER is composed of two basic parts: a set of interactive language specification functions, and a parser. The language specification functions are used to define an application language, a subset of a natural language (e.g., English) that is appropriate for interacting with existing software, such as a DBMS. Using this language specification, the LIFER parser interprets natural language inputs, translating them into appropriate interactions with the application software.

Figure 2 shows simplified example interactions with the LIFER parser using the INLAND language specification. A sequence of complete examples is presented in the Appendix. The user of the system types in a question or command in ordinary English, followed by a carriage return. The LIFER parser then begins processing the input. When analysis is complete, the system types “PARSED!” and invokes database functions (IDA) to respond.

An important feature of the parser is an ability to process elliptical (incomplete) inputs. Thus if the system is asked, as in question 1 of Figure 2,

WHAT IS THE LENGTH OF THE CONSTELLATION,

then the subsequent input

OF THE NAUTILUS

will be interpreted as WHAT IS THE LENGTH OF THE NAUTILUS.

If a user misspells a word, LIFER attempts to correct the error, using the INTERLISP spelling corrector [19]. If the parser cannot account for an input in terms of the application language definition, error messages, such as that produced after question 6, are printed that indicate how much of the input was understood and suggest means of completing the input.

Provision is included in INLAND for interfacing with LIFER’s own language specification functions, making it possible for users to give natural language commands for extending the language itself. In particular, computer-naive users may extend the language accepted by the system by employing easy-to-understand notions such as synonyms and paraphrases. This is illustrated by interactions 7 and 10 in Figure 2.

In using LIFER to define a language for INLAND, we have followed the approach taken by most real-time language processing systems in embedding considerable semantic information in the syntax of the language. Such a language specification is typically called a “semantic grammar.” For example, words like NAUTILUS and DISPLACEMENT are not grouped together into a single <NOUN> category. Rather, NAUTILUS is treated as a <SHIP-NAME> and DISPLACEMENT as an <ATTRIBUTE>. Similarly, very specific sentence patterns such as

WHAT IS THE <ATTRIBUTE> OF <SHIP>

are typically used instead of more general patterns such as

<NOUN-PHRASE> <VERB-PHRASE>.

For each syntactic pattern, the language definer supplies an expression for computing the interpretation of instances of the pattern. INLAND's expressions for sentence-level patterns usually invoke the IDA component to retrieve information from the distributed database.

This method of language specification is easy to understand, easy to use, and, when pursued systematically, allows languages of broad coverage to be defined, as indicated in Figure 1.

To provide a more detailed view of how LIFER has been employed to produce an efficient and effective language processing system, let us examine in detail a highly simplified fragment of the INLAND language specification.

### 3.2 INLAND's Function in Brief

The central notions of how INLAND is constructed may be seen by considering the problem of providing English access to two files of the form

SHIP: (NAM CLASS COMMANDER HOME-PORT HULL# LOC)

CLASS: (CLASSNAME TYPE NATION FUEL LENGTH BEAM DRAFT SPEED)

located on different computers. IDA and FAM together provide levels of insulation from the real situation, so that INLAND need consider only the problem of specifying what subset of the overall database should be queried and what field values within that subset should be returned. IDA will dynamically plan the appropriate joins on the files in the database, and FAM will carry them out. (In the actual LADDER system, the intertwining of multiple files is much more complex than in the current example.)

### 3.3 A Miniature Language Specification

(1) *Productions.* The grammar rules may be viewed as productions of the form

*metasymbol* → *pattern|expression*,

where *metasymbol* is a metasymbol of the application language, *pattern* is a list of symbols and metasymbols in the language, and *expression* is a LISP expression whose value, when computed, is assigned as the value of the metasymbol.<sup>3</sup> The symbol <L.T.G.> (LIFER top grammar) is the highest-level metasymbol of the grammar. The system's answer to complete inputs that match a pattern instantiating <L.T.G.> will be the result of evaluating the associated LISP expression.

For example, the input

PRINT THE LENGTH OF THE KENNEDY

is an instantiation of the sentence-level production

---

<sup>3</sup>In addition to computing values for acceptable applications of the production, the expression may also be used to reject some applications on semantic grounds. Rejection is signaled if the expression returns \*ERROR\* as its value.

1- What is the length of the Constellation  
PARSED!

(LENGTH 1072 feet)

2- of the Nautilus

TRYING ELLIPSIS: WHAT IS THE LENGTH OF THE NAUTILUS  
(LENGTH 319 feet)

3- displacement

TRYING ELLIPSIS: WHAT IS THE DISPLACEMENT OF THE NAUTILUS  
(STANDARD-DISPLACEMENT 4040 tons)

4- length of the fastest American Nuclear sub

TRYING ELLIPSIS: WHAT IS THE LENGTH OF THE FASTEST AMERICAN  
NUCLEAR SUB

(LENGTH 360 feet NAM LOS ANGELES SPEED 30.0 knots)

5- Who commands the Constellation

SPELLING->CONSTELLATION

PARSED!

(COMMANDER CAPT J. ELLISON)

6- Who commands JFK

TRYING ELLIPSIS: ELLIPSIS HAS FAILED

THE PARSER DOES NOT EXPECT THE WORD "JFK" TO FOLLOW

"WHO COMMANDS"

OPTIONS FOR NEXT WORD OR META-SYMBOL ARE:

(SHIP-NAME)

7- Define JFK to be like Kennedy

PARSED!

(JFK is now a synonym for KENNEDY, which is a ship name)

8- Who commands JFK (that is, retry interaction 6)

PARSED!

(COMMANDER CAPT P. MOFFETT)

9- info JFK country

TRYING ELLIPSIS: ELLIPSIS HAS FAILED

(error message omitted)

10- Define "Info JFK country" to be like "what is the country of JFK"

PARSED!

Figure 2: Simplified interactions with LADDER

11- Info JFK country  
PARSED!  
(NATION USA)

12- Info fastest American nuclear submarine speed  
PARSED !  
(SPEED 30.0 knots NAM LOS ANGELES)

13- Nautilus  
TRYING ELLIPSIS: INFO NAUTILUS SPEED  
(SPEED 22 knots)

Fig 2. (continued from previous page)

<L.T.G.> → <PRESENT> THE <ATTRIBUTE> OF <SHIP> |  
(IDA (APPEND <SHIP> <ATTRIBUTE> )).

The input matches the pattern

<PRESENT> THE <ATTRIBUTE> OF <SHIP>,

where <PRESENT> matches PRINT, <ATTRIBUTE> matches LENGTH, and <SHIP> matches the phrase THE KENNEDY. If the semantic values for <SHIP> and <ATTRIBUTE>, computed by means described shortly, are ((NAM EQ JOHN#F.KENNEDY)) and ((? LENGTH)), respectively, then the answer to the question is computed from the expression portion of the production as follows:

(IDA (APPEND <SHIP> <ATTRIBUTE> ))  
→ (IDA (APPEND '((NAM EQ JOHN.#F.KENNEDY))  
'(? LENGTH)))  
→ (IDA '((NAM EQ JOHN.#F.KENNEDY) (? LENGTH))).

(APPEND is a LISP function that appends any number of lists together to form a larger list.) At this point, the IDA component is called with the argument

((NAM EQ JOHN#F.KENNEDY) (? LENGTH))

and the length of the Kennedy is retrieved as

(IDA '((NAM EQ JOHN#F.KENNEDY) (? LENGTH)))  
→ (LENGTH 1072 feet)

In LIFER, productions like the one just shown are defined interactively by issuing commands such as

PD[<L.T.G.>  
<PRESENT> THE <ATTRIBUTE> OF <SHIP> )  
(IDA (APPEND <SHIP> <ATTRIBUTE> ))],

where PD is the production definition function.

(2) *Lexical Entries.* Metasymbols, such as <PRESENT> and <ATTRIBUTE> are often associated with individual words or fixed phrases, which are maintained in

LIFER's lexicons. The LIFER function MS (make set) is used to define a set of words and phrases that may match a particular metasymbol. For example, the call

```
MS[<ATTRIB>
  (CLASS COMMANDER FUEL TYPE NATION LENGTH
   BEAM DRAFT (LOCATION . LOC) (POSITION . LOC)
   (NAME . NAM) (COUNTRY . NATION)
   (NATIONALITY . NATION) ((HOME PORT) . HOME-PORT)
  ((POWER TYPE) . FUEL) ((HULL NUMBER) . HULL#))]
```

is used to define 16 words and fixed phrases that may match the symbol <ATTRIBUTE> (which is used subsequently in defining <ATTRIBUTE> ).

After this call to MS, <ATTRIB> will match the words CLASS, COMMANDER, FUEL, TYPE, NATION, LENGTH, BEAM, and DRAFT. For these words, <ATTRIB> will take as its semantic value the word itself. <ATTRIB> will also match the word LOCATION, but for this match the value of <ATTRIB> will be LOC. Similarly, <ATTRIB> matches POSITION, NAME, COUNTRY, and NATIONALITY, but takes the values LOC, NAM, NATION, and NATION, respectively. <ATTRIB> also matches the two-word phrase HOME PORT, taking HOME-PORT as its value. For the phrase POWER TYPE, the value is FUEL; for HULL NUMBER it is HULL#. (It is assumed that the codes HOME-PORT, HULL#, LOC, and NAM are peculiar to the database and will not occur in natural language inputs.)

(3) *Subgrammars*. Metasymbols may also be defined by production rules. For example, the call

```
PD[<ATTRIBUTE>
  <ATTRIB>
  (LIST (LIST '? <ATTRIB> ))]
```

indicates that an <ATTRIBUTE> may be matched by an <ATTRIB>, viz:

<ATTRIBUTE> → <ATTRIB>.

For this production, the associated expression is

(LIST (LIST '? <ATTRIB> )).

Since the word LENGTH matches <ATTRIB> and cause <ATTRIB> to take LENGTH as its value, the rule above indicates that LENGTH is an instantiation of <ATTRIBUTE>. That is,

<ATTRIBUTE> → <ATTRIB> → LENGTH.

The value assigned to <ATTRIBUTE> when it matches LENGTH is computed by the production's expression as follows:

```
(LIST (LIST '? <ATTRIB> )).
→ (LIST (LIST '? LENGTH)).
→ (LIST '? LENGTH)).
→ ' (? LENGTH)).
```

This fragment of an IDA command requests the value of the LENGTH field. It was used above in answering the question "What is the length of the Kennedy?"

To recognize inputs such as

PRINT THE LENGTH BEAM AND DRAFT OF THE KENNEDY,

the concept of an <ATTRIBUTE> may be generalized<sup>4</sup> by adding two new productions as follows:

```
PD[<ATTRIBUTE>
(<ATTRIB> AND <ATTRIBUTE> )
(CONS (LIST.'? <ATTRIB> ) <ATTRIBUTE> )]

PD[<ATTRIBUTE>
(<ATTRIB> <ATTRIBUTE> )
(CONS (LIST '? <ATTRIB> ) <ATTRIBUTE> )].
```

(CONS is a LISP function that adds an element (in this case the list whose first element is ? and whose second element is the value of <ATTRIB> ) to the front of a list (in this case the value of <ATTRIBUTE> ).) These productions allow the phrase LENGTH BEAM AND DRAFT to be accounted for in terms of the syntax tree of Figure 3.

(4) *Complete Analysis of a Simple Query.* The examples above have indicated how the pattern

<PRESENT> THE ATTRIBUTE OF SHIP

may be defined as a top-level input and how the metasymbol <ATTRIBUTE> may be defined. To complete the analysis of the top-level pattern, consider now the following definitions for <PRESENT> and <SHIP>.

To define <PRESENT>, the function MS may be used:

```
MS[<PRESENT>.
  (PRINT LIST SHOW GIVE ((GIVE ME) . PRINT)
  ((WHAT IS) . PRINT) ((WHAT ARE) . PRINT))].
```

This call allows <PRESENT> to match the words PRINT, LIST, SHOW, a GIVE and the phrases GIVE ME, WHAT IS, and WHAT ARE. The values assigned to <PRESENT>, which might be used, for example, to direct output to the terminal or to a graphics subsystem, are not of interest here.

A <SHIP> may be designated in any one of a number of ways, the simplest being by name. The call

---

<sup>4</sup>The use of two symbols <ATTRIB> and <ATTRIBUTE> could be avoided by letting <ATTRIBUTE> directly match lexical items and by introducing such productions as <ATTRIBUTE> → <ATTRIBUTE> AND <ATTRIBUTE>. Unfortunately, the collapse of the two symbols into one results in both ambiguity and left recursion. LIFER recognizes only one of the ambiguous interpretations. Left recursion can be tolerated by special mechanisms in LIFER's top-down left-to-right parser, but only at a considerable increase in parsing time.

```
PD[<SHIP>
  (<SHIP-NAME> )
  (LIST (LIST 'NAM 'EQ <SHIP-NAME> ))]
```

causes **<SHIP>** to match a **<SHIP-NAME>** and to take as its value an IDA Command fragment restricting the value of the NAM field to be EQ (equal) to the particular name. **<SHIP-NAME>** may be defined by MS:

```
MS[<SHIP-NAME>
  (CONSTELLATION NAUTILUS
  (KENNEDY . JOHN#F.KENNEDY)
  ((JOHN F . KENNEDY) . JOHN#F.KENNEDY) etc.)].
```

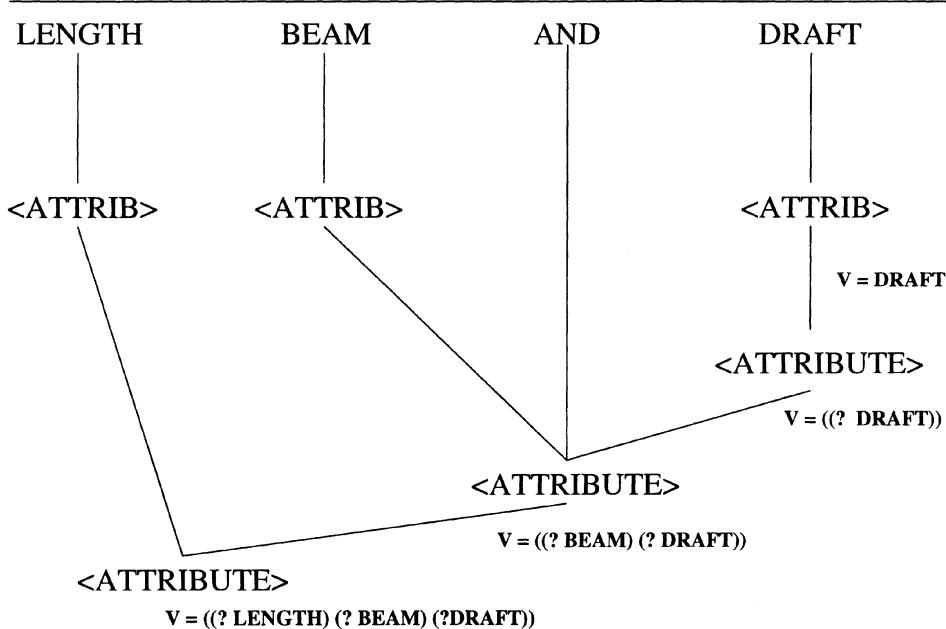


Figure 3: **<ATTRIBUTE>** syntax tree

For an actual database, this list is, of course, much more extensive. To allow the optional use of "the" before the name of a ship, a supplementary production for **<SHIP>** may be defined:

```
PD[<SHIP>
  (THE <SHIP> )
  <SHIP>].
```

With these definitions, LIFER has been given all the information needed to process a small class of sentence-level inputs. For example, the complete analysis of the input

## WHAT IS THE LENGTH OF THE KENNEDY

is shown in the syntax tree of Figure 4. note how the query given to ID. as generated by combining fragments from (SHIP) and (ATTRIBUTE).

From the definitions for complete inputs defined above, LIFER can infer how to process incomplete inputs in context. For example, having just parsed the input

WHAT IS THE LENGTH OF THE KENNEDY,

the system may, without additional knowledge, also handle the following sequence of incomplete inputs:

BEAM

(i.e. what is the beam of the Kennedy)

HOME PORT AND CLASS

(i.e. what is the home port and class of the Kennedy)

NAUTILUS

(i.e. what is the home port and class of the Nautilus).

The method by which these incomplete inputs are processed is discussed below Other inputs that the rules defined thus far will accept include:

GIVE ME THE POSITION OF THE NAUTILUS

PRINT THE HULL NUMBER AND POWER TYPE OF CONSTELLATION

SHOW THE COMMANDER COUNTRY AND TYPE OF THE JOHN F. KENNEDY.

### 3.4 Some Generalizations

The tiny fragment of language defined above already allows English access to most of the fields in the example database, given the name of a ship. This fragment may be expanded easily along many dimensions.

(1) *Generalizing <SHIP>*. Generalizing <SHIP> provides one of the most fruitful expansions. Naval ships are divided into major sets called classes. For example, the Constellation is in the Kitty Hawk class. Sometimes users will wish to ask questions about all ships of a particular class; for example, HOW LONG ARE KITTY HAWK CLASS SHIPS. To do this, the language may be extended by the call

```
PD[<SHIP>
  (<CLASS> CLASS SHIP)
  (LIST (LIST 'CLASS 'EQ <CLASS> ))],
```

Where <CLASS> is defined to match class names and takes their database designations as values.<sup>5</sup> After this extension, the system will accept such inputs as

---

<sup>5</sup>To simplify the language definition, a language builder may supply LIFER with a preprocessor that does certain kinds of morphological transformations. For example, plural nouns such as SHIPS may be converted to the singular SHIP plus the pluralizing suffix -S. Or, as is assumed here, the suffix may simply be discarded.

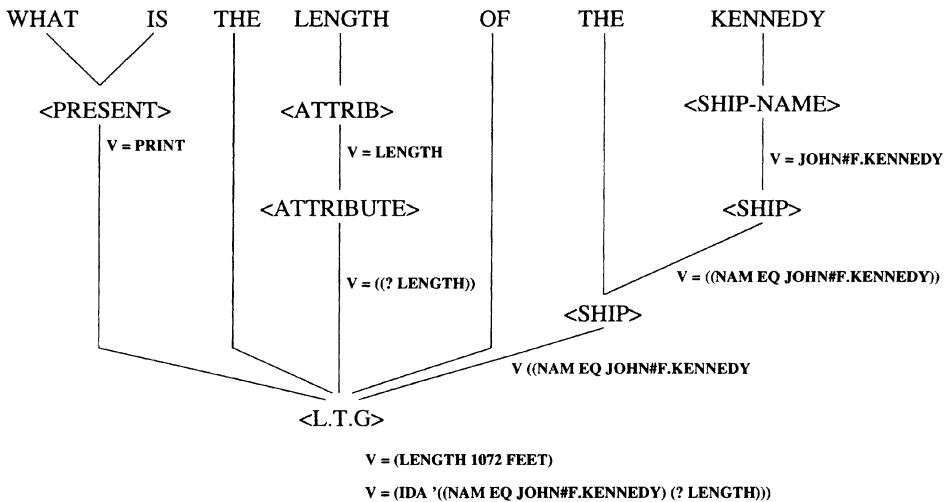


Figure 4: Syntax tree for a complete question

PRINT THE LENGTH OF KITTY HAWK CLASS SHIPS.

A **<SHIP>** might also match a general category such as CARRIERS, CRUISERS or MERCHANT SHIPS. Such categories may usually be defined in terms of the TYPE field in the database. For example, CARRIERS are of type CVA, CVAN or CVS. OILERS are AO or AOR. **<CATEGORY>** might be defined by

```
MS[<CATEGORY>
  ((CARRIER.((TYPE EQ CV)
             OR (TYPE EQ CVAN)
             OR (TYPE EQ CVS)))
   (OILER.(TYPE EQ AO)
          OR (TYPE EQ AOR)))
  etc.)].
```

A new production for **<SHIP>** may then be added such as

```
PD[<SHIP>
  (<CATEGORY> )
  (LIST <CATEGORY> )].
```

With this production, the command

PRINT THE LOCATION OF CARRIERS

will be accepted.

Modifiers such as AMERICAN, NUCLEAR, and CONVENTIONAL are also very useful; for example,

MS[<MOD>  
((AMERICAN . (NATION EQ US))  
(NUCLEAR . (FUEL EQ NUCLEAR))  
(CONVENTIONAL . (FUEL EQ DIESEL))  
etc.)].

By adding

PD [<SHIP>  
<MOD> <SHIP> )  
(CONS <MOD> <SHIP> )],

the system will then process inputs such as

GIVE ME THE POSITION OF THE AMERICAN NUCLEAR CARRIERS.

Superlative modifiers, such as FASTEST and SHORTEST, may be defined:

MS[<MOD>  
((FASTEST . (\* MAX SPEED))  
(SLOWEST . (\* MIN SPEED))  
(LONGEST . (\* MAX LENGTH))  
etc.)]

Then the system will accept inputs such as

GIVE ME THE NAME AND LOCATION OF THE FASTEST AMERICAN OILERS.

This would translate into the IDA call

(IDA'((\*MAX SPEED) (NATION EQ US)  
((TYPE EQ AO) OR (TYPE EQ AOR))  
(? NAM) (? LOC))).

(2) *Generalizing<L.T.G>*. New sentence-level productions, defined in terms of the more primitive metasymbols already described to LIFER, also greatly extend the range of language accepted. For example,

PD [<L.T.G>  
<PRESENT> <SHIP> )  
(IDA (CONS'(? NAM) <SHIP> ))]

allows inputs such as

WHAT ARE THE FASTEST NUCLEAR SUBMARINES  
PRINT THE CARRIERS

and

GIVE ME THE KITTY HAWK CLASS SHIPS.

As another example,

```
PD[<L.T.G>
  (WHO COMMANDS THE <SHIP> )
  (IDA (CONS '(? COMMANDER) <SHIP> ))]
```

allows the input

WHO COMMANDS THE KENNEDY.

(3) *Calculated Answers.* Sometimes a database does not contain the information needed to answer a question directly, but nevertheless contains information that may be used as input to a procedure that can compute the answer from more primitive data. For example, the distance between two ships is not directly available in the example database, although position data is. Suppose the function STEAMING TIME can take speed and position information returned by IDA and calculate the time for the first ship to travel to the position of the second ship. Then, after defining <SHIP2> like <SHIP>,

```
PD [<SHIP2>
  (<SHIP> )
  <SHIP>],
```

a new top-level production may be defined as follows:

```
PD [<L.T.G>
  (HOW MANY HOURS IS <SHIP> FROM <SHIP2> )
  (STEAMING-TIME
  (IDA (APPEND '((? SPEED)(? LOC)) <SHIP> ))
  (IDA (CONS ' (? LOC) <SHIP2> )))].
```

This production allows such queries as

HOW MANY HOURS IS KENNEDY FROM THE CONSTELLATION.

### 3.5 Extending the Lexicon with Predicates

In certain instances, it is impractical to use the MS function to explicitly list all of the symbols that might match some metasymbol. For example, if the metasymbol <NUMBER> is to match any number, then MS is of little value. For such cases, LIFER allows a metasymbol to be associated with a predicate function. The metasymbol will match any symbol for which the predicate returns a non-NIL value. When such a match occurs, the metasymbol will take as its semantic value the response returned by the application of the predicate. To define a metasymbol in terms of a predicate, the function MP (make predicate) is used. For example,

```
MP[<NUMBER> NUMBERP]
```

defines <NUMBER> to match any symbol for which LISP predicate function NUMBERP returns a non-NIL value. When applied to numbers, NUMBERP returns the number itself. When applied to anything else, it returns NIL.

As the following questions indicate, <NUMBER> has many uses in the example database:

WHAT CARRIERS HAVE LENGTHS GREATER THAN 1000 FEET  
HOW FAR IS CONSTELLATION FROM 40 DEGREES NORTH 6 DEGREES EAST  
WHAT SHIPS ARE WITHIN 100 MILES OF KENNEDY.

As the size of the lexicon becomes large, the predicate feature may be used to push certain large classes of words out of the natural language system and into the database itself. For example, <SHIP-NAME> could be defined in terms of a predicate that accesses the NAM field of the database. (This would slow the parsing operation, of course, and spelling correction could not be performed easily.)

### 3.6 Accepting Metalanguage Inputs

(1) *Interrogating the Language System.* It is possible to define input patterns that make reference to the LIFER package itself. For example, LIFER contains a function called SYMBOL.INFO which takes a metasymbol as its argument and its lexical items, patterns, and predicates that may be used to match the symbol. The interface builder may incorporate this function in response expressions as in

```
PD[<L.T.G>
  (HOW IS <SYMBOL> USED)
  (SYMBOL.INFO <SYMBOL> )]
```

After this call to PD,<sup>6</sup> a user might ask the metaquestion

HOW IS <SHIP> USED

and receive the reply

<SHIP> MAY BE ANY SEQUENCE OF WORDS FOLLOWING ONE OF THE PATTERNS:

```
<SHIP> → <SHIP-NAME>
          THE <SHIP>
          <CLASS> CLASS SHIP
          <MOD> <SHIP>.
```

Using other system interrogation functions, it is possible to provide in an application language for such inputs as

```
PRINT THE GRAMMAR ON FILE APP.GRAM
DISPLAY THE PRODUCTIONS EXPANDING <SHIP>
SHOW LEXICAL ENTRIES FOR CATEGORY <SHIP-NAME>
```

---

<sup>6</sup>As specified more fully in (Hendrix 1977a), the metasymbol <SYMBOL> may itself be defined by function MP, using a predicate that sees whether its argument is included in the list of defined metasymbols. Being so defined, <SYMBOL> can even match itself so that the input HOW IS <SYMBOL> USED may be parsed and answered.

WHAT PREDICATE DEFINES <NUMBER>  
DRAW THE SYNTAX TREE FOR THE LAST INPUT  
HOW WOULD YOU PARSE "HOW FAST IS KENNEDY"  
IN WHAT PRODUCTIONS DOES <SHIP> APPEAR ON THE RIGHT.

Metaquestions requesting general information about the system, such as

WHAT KIND OF INFORMATION DO YOU KNOW ABOUT  
WHAT'S IN THE DATABASE  
HELP.

may easily be included in the application language. Top-level response expressions or such inputs may simply return canned explanation texts. With more sophistication, the expressions might access a semantic schema of the database to help formulate an up-to-date reply.

(2) *Personalizing the Application Language.* LIFER contains a function called SYNONYM that allows a new word to be defined as having the same meaning as model word that is already known to LIFER. Using this function, an interface builder may introduce structures into the application language that allows users to define their own synonyms at run time. In particular,

```
PD[<L.T.G>
  (DEFINE <NEW-WORD> LIKE <OLD-WORD> )
  (SYNONYM <NEW-WORD> <OLD-WORD> )]
```

allows the parser to accept inputs such as

```
DEFINE JFK LIKE KENNEDY.
```

The symbols <NEW WORD> and <OLD-WORD> are defined by predicates that will match any word. SYNONYM works by copying lexical information from <OLD-WORD> to <NEW-WORD>.

LIFER also contains a function called PARAPHRASE that allows a new sequence of words to be defined as having the same meaning as a model sequence of words that the parser already accepts as a complete sentence. Using function PARAPHRASE in a response expression, the interface builder may extend the grammar by

```
PD [<L.T.G>
  (LET <NEW-SEQUENCE> BE A PARAPHRASE OF
    <OLD-SENTENCE> )
  (PARAPHRASE <NEW-SEQUENCE> <OLD-SENTENCE> )]
```

where <NEW-SEQUENCE> matches any sequence of words and returns a list of the matched words as its value, and <OLD-SENTENCE> matches any sequence of words currently accepted as a sentence in the application language.

This new rule allows computer-naive users to personalize the syntactic constructions understood by the system at run time. For example, the user might say

```
LET "REPORT ON KENNEDY" BE A PARAPHRASE OF
```

“PRINT THE LOCATION AND COMMANDER OF KENNEDY”.

The expression associated with the top-level production that matches this input sentence calls upon the paraphraser. Given the language definition defined above, LIFER then automatically adds the new production

<L.T.G> → REPORT ON <SHIP>

to the system, with an appropriate response expression. This new, user-defined production will allow the system to accept such new inputs as

REPORT ON THE KENNEDY  
REPORT ON OILERS  
REPORT ON THE FASTEST AMERICAN SUBMARINES.

LIFER’s methods for learning paraphrases are discussed below.

### 3.7 Extendibility

The preceding subsections have indicated how a few simple notions may be drawn together to create a small interface. But can the same notions be used to create much more sophisticated systems? Until our recent experience, we would have joined others in answering, “Not likely”. Long before reaching an acceptable level of performance, previous language systems, including our own, have generally grown so complex and unwieldy that further extension has been stifled.

In designing LIFER, much attention has been given to the problem of supplying interface builders with an environment supporting the incremental development of relatively broad interfaces. All LIFER functions are interactive. Parsing and language specification tasks may be intermixed, allowing interface builders to operate in a rapid, extend-and-test mode. Transition trees, which are an efficient representation for the parser to work with, are automatically produced from productions, which we have found to be an efficient representation for interface builders to work with. The system contains a grammar editor and numerous special functions for answering questions about the structure of the language definition and for tracing and debugging a grammar. Details concerning these and other features of LIFER are specified more fully in the LIFER Manual (Hendrix 1977a).

We believe that the support features of LIFER have enabled us to give the INLAND language broader coverage than previous systems. Unfortunately, We know of no adequate measure of “breadth of coverage”. However, some feeling for the types of inputs accepted by LADDER may be gained by considering a sample of acceptable inputs, such as those shown in Figure 1.

## 4 The Transition Tree Parser

The LIFER parser is a top-down, left-to-right parser based on a simplification of the augmented transition network (ATN) system developed by Woods [23]. Rather than use

true ATNs, LIFER works with transition trees. If  $\langle L.T.G \rangle$  is defined by the productions

- $\langle L.T.G \rangle \rightarrow \langle PRESENT \rangle \text{ THE } \langle ATTRIBUTE \rangle \text{ OF}$
- $\quad \quad \quad \langle SHIP \rangle \mid e1$
- $\rightarrow \langle PRESENT \rangle \langle SHIP'S \rangle \langle ATTRIBUTE \rangle \mid e2$
- $\rightarrow \text{HOW MANY } \langle SHIP \rangle \text{ ARE THERE} \mid e3$
- $\rightarrow \text{HOW MANY } \langle SHIP \rangle \text{ ARE THERE WITH}$
- $\quad \quad \quad \langle PROPERTY \rangle \mid e4$ ,

syntax tree of Figure 5 would be constructed for use by the parser.

Starting at the box labeled  $\langle L.T.G \rangle$ , the parser attempts (nondeterministically) to move toward the response expressions on the right. At each step, the parser may move to the right on a branch if the left part of the remaining portion of the input can be matched by the symbol on the branch. Literal words on a branch can be matched only by themselves. A metasymbol, such as  $\langle PRESENT \rangle$ , may be matched by a lexical item in the associated set created by MS. Or it may be matched by the predicate, if any, that has been defined for the metasymbol. Or it may be matched by successfully transversing some branch of the transition tree that encodes the productions expanding the metasymbol.

At the top level, if the parser reaches a response expression as a result of accounting for the last word of an input, then a top-level match for the input has been found and the response expression is evaluated to compute a response.

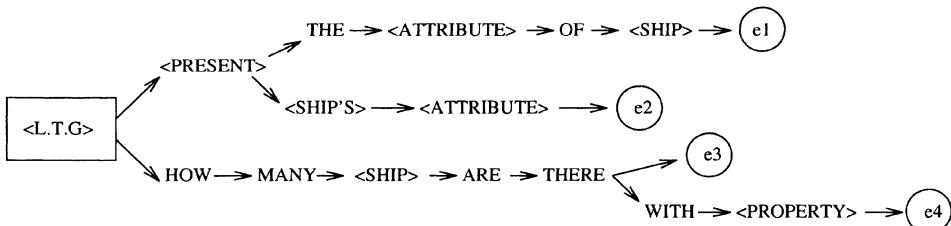


Figure 5: Transition tree

## 5 Implementation of Special Lifer Features

This section presents an overview of LIFER's implementation of the spelling corrector, elliptical processor, and paraphraser.

### 5.1 Implementation of Spelling Correction

Each time LIFER's left-to-right, ATN parser discovers that it can no longer follow transitions along the current path, it records the failure on a failpoint list. Each entry on this list indicates the state of the system when the failure occurred (i.e. the position in the transition net and the values of various stacks and registers and the current position in the input string. Local ambiguities and false paths make it quite normal for many failpoints to be noted even when a perfectly acceptable input is processed.

If a complete parse is found for an input, the failpoints are ignored. But if a input cannot be parsed, the list of failpoints is used by the spelling corrector, which selects those failpoints associated with the rightmost position in the input at which failpoints were recorded. It is assumed that failpoints occurring to the left were not caused by spelling errors, since some transitions using the words at those positions must have been successful for there to be failpoints to their right.<sup>7</sup>

The spelling corrector further restricts the rightmost failpoints by looking for cases in which a rightmost failpoint G is dominated by another rightmost failpoint F. G is dominated by F if G is a failpoint at the beginning of a subordinate transition tree that was reached in an attempt to expand F.

Working with the rightmost dominating failpoints, the spelling corrector finds all categories of words that would be valid at the point where the suspected misspelling occurred. This typically requires an exploration of subgrammars. Using the INTERLISP spelling corrector, the word of the input string associated with the rightmost failpoints is compared with the words of the categories just found. If the misspelled word is sufficiently similar to any of these lexical items, the closest match is substituted. Failpoints associated with lexical categories that include the new word are then sequentially restarted until one leads to a successful parse. (This may produce more spelling corrections further to the right.) If all restarts with the new word fail, other close lexical items are substituted for the misspelled word. If these also fail, LIFER prints an error message.

## 5.2 Implementation of Ellipsis

LIFER's mechanism for treating elliptical inputs presumes that the application language is defined by a semantic grammar so that a considerable amount of semantic information is encoded in the syntactic categories. Thus similar syntactic constructions are expected to be similar semantically. LIFER's treatment of ellipsis is based on this notion of similarity. During elliptical processing, LIFER is prepared to accept any string of words that is syntactically analogous to any contiguous substring of words in the last input. (If the last input was elliptical, its expansion into a complete sentence is used.)

LIFER's concept of analogy appeals to the syntax tree of the last input that was successfully analyzed by the system. For any contiguous substring of words in the last input, an "analogy pattern" may be defined by an abstraction process that works backward through the old syntax tree from the words of the substring toward the root. Whenever the syntax tree shows a portion of the substring to be a complete expansion of a syntactic category, the category name is substituted for that portion. The analogy pattern is the final result after all such substitutions. For example, consider how an analogy pattern may be found for the substring

OF SANTA INEZ,

---

<sup>7</sup>This heuristic can cause LIFER to fail to find and correct certain errors. For example, if the user types CRAFT for DRAFT in WHAT DRAFT DOES THE ROARK HAVE, the spelling error will not be caught since a sentence such as WHAT CRAFT ARE NEAR ROARK would account for the initial sequence WHAT CRAFT. This is traded off against faster processing for the majority of errors.

using the syntax tree shown in Figure 6 for a previous input, WHAT IS THE LENGTH OF SANTA INEZ.

Note that the syntax tree used in Figure 6 reflects production rules similar to those defined previously, but introduces a new metasymbol, <ITEM>, to add more substance to the discussion. Since the SANTA INEZ portion of the substring is a complete expansion of <SHIP-NAME>, the substring is rewritten as OF <SHIP-NAME>. Similarly, since <SHIP> expands to <SHIP-NAME>, the substring is rewritten as OF <SHIP>. Since no other portions of the substring are complete expansions of other syntactic categories in the tree, the process stops and OF <SHIP> is accepted as the most general analogy pattern. If the current input matches this analogy pattern, LIFER will accept it as a legitimate elliptical input. For example, the analogy pattern OF <SHIP>, extracted from the last input, may be used to match such current elliptical inputs as

OF THE KENNEDY  
OF THE FASTEST NUCLEAR CARRIER

and

OF KITTY HAWK CLASS SHIPS.

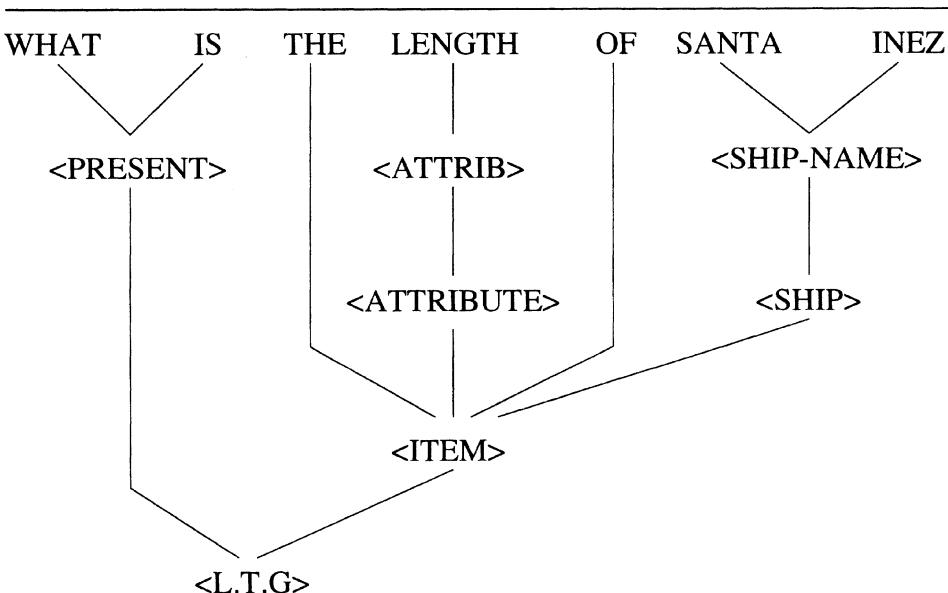


Figure 6: Syntax tree

---

Note that the expansion of <SHIP> need not parallel its expansion in the old input

that originated the analogy pattern. For example, OF KITTY HAWK CLASS SHIPS is not matched by expanding <SHIP> to <SHIP-NAME> but by expanding <SHIP> to <CLASS> CLASS SHIP.

To compute responses for elliptical inputs matching OF <SHIP>, LIFER works its way back through the old syntax tree from the common parent of OF <SHIP> toward the root. First, the routine for computing the value of an <ITEM> from constituents of the production

<ITEM> → THE <ATTRIBUTE> OF <SHIPS>

is invoked, using the new value of <SHIP> (which appeared in the current elliptical input) and the old value of <ATTRIBUTE> from the last sentence. Then, using the newly computed value for <ITEM> and the old value for <PRESENT>, a new value is similarly computed for <L.T.G>, the root of the syntax tree. Some other substrings with their associated analogy patterns are shown below along with possible new elliptical inputs matching the patterns:

substring:	THE LENGTH
pattern:	THE <ATTRIBUTE>
a match:	THE BEAM AND DRAFT
substring:	LENGTH OF SANTA INEZ
pattern:	<ATTRIBUTE> OF <SHIP>
a match:	HOME PORTS OF AMERICAN CARRIERS
substring:	WHAT IS THE LENGTH
pattern:	<PRESENT> THE <ATTRIBUTE>
a match:	PRINT THE NATIONALITY
substring:	WHAT IS THE LENGTH OF SANTA INEZ
pattern:	<L.T.G>
a match:	[any complete sentence]

For purposes of efficiency, LIFER's elliptical routines have been coded in such a way that the actual generation of analogy patterns is avoided.<sup>8</sup> Nevertheless, the effect is conceptually equivalent to attempting parses based on the analogy pattern of each of the contiguous substrings of the last input.

### 5.3 Implementation of Paraphrase

LIFER's paraphrase mechanism also takes advantage of semantically-oriented syntactic categories and makes use of syntax trees. In the typical case, the paraphraser is given a model sentence, which the system can already understand, and paraphrase. The paraphraser's general strategy is to analyze the model sentence and then look for similar structures in the paraphrase string. In particular, the paraphraser invokes the parser to produce a syntax tree of the model. Using this tree, the paraphraser determines all proper subphrases of the model, i.e. all substrings that are complete expansions of one

---

<sup>8</sup>See (Hendrix 1977b) for details of the algorithm.

of the syntactic categories listed in the tree. Any of these model subphrases that also appear in the paraphrase string are assumed to play the same role in the paraphrase as in the model itself. Thus the semantically-oriented syntactic categories that account for these subphrases in the model are reused to account for the corresponding subphrases of the paraphrase. Moreover, the relationship between the syntactic categories that is expressed in the syntax tree of the model forms a basis for establishing the relationship between the corresponding syntactic units inferred for the paraphrase.

(1) *Defining a Paraphrase Production.* To find correspondences between the model and the paraphrase, the subphrases of the model are first sorted. Longer phrases have preference over shorter phrases, and for two phrases of the same length, the leftmost is taken first. For example, the sorted phrases for the tree of Figure 6 are

- |                |                          |
|----------------|--------------------------|
| 1. <ITEM>      | THE LENGTH OF SANTA INEZ |
| 2. <PRESENT>   | WHAT IS                  |
| 3. <SHIP-NAME> | SANTA INEZ –not used     |
| 4. <SHIP>      | SANTA INEZ               |
| 5. <ATTRIB>    | LENGTH –not used         |
| 6. <ATTRIBUTE> | LENGTH.                  |

Because the syntax tree indicates  $\langle \text{SHIP} \rangle \rightarrow \langle \text{SHIP-NAME} \rangle \rightarrow \text{SANTA INEZ}$ , both  $\langle \text{SHIP-NAME} \rangle$  and  $\langle \text{SHIP} \rangle$  account for the same subphrase. For such cases, only the most general syntactic category ( $\langle \text{SHIP} \rangle$ ) is considered. The category  $\langle \text{ATTRIB} \rangle$  is similarly dropped.

Beginning with the first (longest) subphrase, the subphrases are matched against sequences of words in the paraphrase string. (If a subphrase matches two sequences of words, only the leftmost match is used.) The longer subphrases are given preference since matches for them will lead to generalizations incorporating matches for the shorter phrases contained within them. Whenever a match is found, the syntactic category associated with the subphrase is substituted for the matching word sequence in the paraphrase. This process continues until matches have been attempted for all subphrases.

For example, suppose the paraphrase proposed for the question of Figure 6 is

FOR SANTA INEZ GIVE ME THE LENGTH.

Subphrases 1 and 2, listed above, do not match substrings in this paraphrase. Subphrase 3 is not considered, since it is dominated by subphrase 4. Subphrase 4 does match a sequence of words in the paraphrase string. Substituting the associated category name for the word sequence yields a new paraphrase string:

FOR <SHIP> GIVE ME THE LENGTH.

Subphrase 5 is not considered, but subphrase 6 matches a sequence of words in the updated paraphrase string. The associated substitution yields

FOR <SHIP> GIVE ME THE <ATTRIBUTE>.

Since there are no more subphrases to try, the structure

<L.T.G.> → FOR <SHIP> GIVE ME THE <ATTRIBUTE>

is created as a new production to account for the paraphrase and for similar inputs such as  
FOR THE FASTEST AMERICAN SUB GIVE ME THE POSITION AND HOME PORT.

(2) *Defining a Response Expression for the Paraphrase Production.* A new semantic response expression indicating how to respond to inputs matching this paraphrase production is programmed automatically from information in the syntax tree of the model. In particular, the syntax tree indicates which productions were used in the model to expand various syntactic categories. Associated with each of these productions is the corresponding response expression for computing the interpretation of the subphrase from subphrase constituents. The paraphraser reuses selected response expressions of the model to create a new expression for the paraphrase production. The evaluation of this new expression produces the same effect that would be produced if the expressions of the model were reevaluated. Metasymbols that appear in both the paraphrase production and the model remain as variables in the new response expression. Those symbols of the model that do not appear in the paraphrase production are replaced in the expression by the constant values to which they were assigned in the model.

## 6 Discussion

As implied by Figure 1 and the examples given in the Appendix, the INLAND system is a habitable, rather robust, real-time interface to a large database and is fully capable of successfully accepting natural language inputs from inexperienced users. In the preceding sections, we have indicated some of the key techniques used in creating this system. We now seek to place our previous remarks in perspective by considering some of the limitations of the system, the roles played by the nature of our task and the tools we built in developing the system, and some of the similarities and differences between other systems and our own.

### 6.1 Limitations

In considering the limitations of our work, the reader should distinguish between limitations in the current INLAND grammar and limitations in the underlying LIFER system.

#### (1) Syntactic Limitations

(a) *The Class of Languages Covered by LIFER.* Consider the set of sentences that LIFER can accept. Because in the worst case a special top-level production may be defined in LIFER to cover any (finite-length) sentence that an interface builder may wish to include in the application language, it is impossible to exhibit a single sentence that the LIFER parser cannot be made to accept. Therefore, the only meaningful questions

concerning syntactic limitations of LIFER must relate to LIFER's ability to use limited memory in covering infinite or large finite set of sentences.

LIFER application languages are specified by augmented context-free<sup>9</sup> grammars. Each rule in the grammar, as discussed previously, includes a context-free production, plus an arbitrarily complex response expression, which is the "augmentation." Although a purely context-free system would severely restrict the set of (nonfinite) languages that LIFER could accept, the use of augmentation gives the LIFER parser the power of a Turing machine. The critical question is whether or not the context-free productions and their more powerful augmentations can be made to support one another in meaningful ways.

To see the interplay between augmentation and context-free rules in the recognition of a classic example of non-context-free languages, consider the language composed of one or more X's followed by an equal number of Y's followed by an equal number of Z's. Let  $\langle x \rangle$  be defined as

$$\begin{array}{lcl} \langle x \rangle & \rightarrow & X \\ \langle x \rangle & \rightarrow & X \langle x \rangle \quad | \quad (PLUS \ 1 \ \langle x \rangle) \end{array}$$

Thus  $\langle x \rangle$  matches an arbitrary sequence of X's and takes as its value the number of X's in the string. Similar definitions may be made for  $\langle y \rangle$  and  $\langle z \rangle$ . A top-level sentence may be defined by the pattern  $\langle x \rangle \langle y \rangle \langle z \rangle$ , but the augmentation must check to see that the numeric values assigned to the metasymbols are all equal. If they are equal, the augmentation expression returns some appropriate response. But if they are unequal, the expression returns the special symbol \*ERROR\*, which the LIFER parser traps as a "semantic" (as opposed to syntactic) rejection. The Turing machine power of LIFER is illustrated by the following trivial grammar:

$$\begin{array}{ll} \langle PRE-SENTENCE \rangle & \rightarrow \langle WORD \rangle \mid (\text{LIST } \langle WORD \rangle) \\ & \rightarrow \langle WORD \rangle \langle PRE-SENTENCE \rangle \\ & \quad | \quad (\text{CONS } \langle WORD \rangle \\ & \quad \quad \langle PRE-SENTENCE \rangle) \\ \\ \langle SENTENCE \rangle & \rightarrow \langle PRE-SENTENCE \rangle \\ & \quad | \quad (\text{TMPARSE } \langle PRE-SENTENCE \rangle) \end{array}$$

This grammar simply collects all of the words of the input into a list which is then passed to function TMPARSE, a parser of Turing machine power. In this extreme case, the LIFER parser makes virtually no use of the context-free productions, but relies exclusively on the augmentation. LIFER is best used in the middle ground between this extreme and a purely context-free system.

In other words, the class of languages for which LIFER was designed may be characterized as those allowing much of their structure to be defined by context-free rules but requiring occasional augmentation. It has been our experience that much of the subset of English used for asking questions about a command and control database falls in this class. However, we have not considered certain complex types of transformations which will be discussed in the next subsection.

---

<sup>9</sup>See (Hopcroft and Ullman 1969) for definitions of terms such as "context free" and "context sensitive."

(b) *Troublesome Syntactic Phenomena.* English speakers and writers often omit from a sentence a series of words that do not form a complete syntactic unit. For example, consider the following family of conjunctive sentences:

1. WHAT LAFAYETTE AND WASHINGTON CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR
2. WHAT LAFAYETTE CLASS AND WASHINGTON CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR
3. WHAT LAFAYETTE CLASS SUBS AND KITTY HAWK CLASS CARRIERS IN THE ATLANTIC ARE WITHIN 500 MILES OF GIBRALTAR
4. WHAT LAFAYETTE CLASS SUBS IN AND PORTS ON THE ATLANTIC ARE WITHIN MILES OF GIBRALTAR
5. WHAT LAFAYETTE CLASS SUBS IN THE ATLANTIC AND KITTY HAWK CLASS CARRIERS IN THE MEDITERRANEAN SOON WILL BE WITHIN 500 MILES OF GIBRALTAR.

Sentence (1) omits the fragment CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR from the “complete” question WHAT LAFAYETTE CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR AND WHAT WASHINGTON CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR. Note that the omitted fragment does not correspond to any well-formed syntactic unit, but begins in the middle of the noun phrase WHAT LAFAYETTE CLASS SUBS and continues to its right. Moreover, the fragment of the noun phrase that is left behind, namely, WHAT LAFAYETTE, is not likely to be a well-formed syntactic unit, because one would expect to have WHAT combine with LAFAYETTE CLASS-SUBS rather than have WHAT-LAFAYETTE combine with CLASS SUBS. As the family of sentences above illustrates, the omission of words, signaled by the conjunction AND, may be moved to the right through the sentence one word at a time, slicing up the well formed syntactic units at arbitrary positions. INLAND has no difficulty in accepting either conjunctions or disjunctions of well-formed syntactic categories, but LIFER provides no general mechanism for dealing with omissions that slice through categories at arbitrary points.

In the SYSCONJ facility of Woods (Woods 1973), special mechanisms for handling large (but not exhaustive) class of conjunction constructions were built into the parser. Roughly, when SYSCONJ encounters the conjunction “AND” in an input X AND Y, it nondeterministically attempts to break both X and Y into three (possibly empty) parts X<sub>1</sub>-X<sub>2</sub>-X<sub>3</sub> and Y<sub>1</sub>-Y<sub>2</sub>-Y<sub>3</sub>, such that X<sub>1</sub>-X<sub>2</sub>-X<sub>3</sub>-Y<sub>2</sub>-Y<sub>3</sub> and X<sub>1</sub>-X<sub>2</sub>-Y<sub>1</sub>-Y<sub>2</sub>-Y<sub>3</sub> parse as sentences with the same basic syntactic structure. In particular, X<sub>2</sub>-X<sub>3</sub>-Y<sub>2</sub> and X<sub>2</sub>-Y<sub>1</sub>-Y<sub>2</sub> must be expansions of the same metasymbol. The effect of SYSCONJ is to transform X<sub>1</sub>-X<sub>2</sub>-X<sub>3</sub>-AND-Y<sub>1</sub>-Y<sub>2</sub>-Y<sub>3</sub> into X<sub>1</sub>-X<sub>2</sub>-X<sub>3</sub>-Y<sub>2</sub>-Y<sub>3</sub> and X<sub>1</sub>-X<sub>2</sub>-Y<sub>1</sub>-Y<sub>2</sub>-Y<sub>3</sub>.

For example,

WHAT LAFAYETTE AND WASHINGTON CLASS SUBS ARE THERE

may be analyzed as

<u>WHAT</u>	<u>empty</u>	<u>LAFAYETTE</u>	<u>AND</u>
X1	X2	X3	and
<u>WASHINGTON</u>	<u>CLASS SUBS</u>	<u>ARE THERE</u>	
Y1	Y2	Y3	

Both X1-X2-X3-Y2-Y3 (WHAT LAFAYETTE CLASS SUBS ARE THERE) and X1-X2-Y1-Y2-Y3 (WHAT WASHINGTON CLASS SUBS ARE THERE) are parsed by what would correspond in INLAND to the sentence-level production

<L.T.G> → WHAT <SHIP> ARE THERE

and both X2-X3-Y2 (LAFAYETTE CLASS SUBS) and X2-Y1-Y2 (WASHINGTON CLASS SUBS) are expansions of the same metasymbol, <SHIP>. In effect, the original input is transformed into WHAT LAFAYETTE CLASS SUBS ARE THERE and WHAT WASHINGTON CLASS SUBS ARE THERE. Handling conjunctions is just one example of the general need to perform transformations at parse time. A similar phenomenon occurs with comparative clauses, but much more is omitted and transformed. For example,

THE KITTY HAWK CARRIES MORE MEN THAN THE WASHINGTON

may be viewed as a transformed and condensed form of

THE KITTY HAWK CARRIES X-MANY MEN AND  
THE WASHINGTON CARRIES Y-MANY MEN AND  
X IS MORE THAN Y.

For further discussion of this subject, see Paxton (Paxton 1977).

(c) YES/NO *Questions*. A limitation of INLAND, although not of LIFER, is that few YES/NO questions are covered. The reason for this is pragmatic— INLAND users do not ask them. Upon reflection, the motivation for this is clear— WH questions (i.e. questions asking who, what, where, when, or how) produce more information for the questioner at a lower cost. A user might ask

IS THE KENNEDY 1000 FEET LONG,

but it is shorter to ask

HOW LONG IS THE KENNEDY,

and if the answer to the first question is NO (and if the system is so inconsiderate as to not indicate the correct length), then the user may still have to ask for the length.

Creating a grammar for YES/NO questions is easy enough. For example,

```
PD[<L.T.G>
  (IS <NUMBER> <UNIT> THE <ATTRIB> OF <SHIP> )
  (YESNO.NUM.ATT <NUMBER> <UNIT> <ATTRIB>
   <SHIP> )]
```

might be used to allow the input

IS 1000 FEET THE LENGTH OF THE KENNEDY.

Function YESNO.NUM.ATT finds the <ATTRIB> of <SHIP> using IDA. Knowing the units in which the database stores values of <ATTRIB>, YESNO.NUM.ATT converts the returned answer into the units specified by <UNIT> and compares the converted value to <NUMBER>. If the units are valid and the numbers match, YES is returned; otherwise NO is returned, and the correct answer, as computed by IDA, is printed.

(d) *Assertions*. INLAND was designed for retrieval and therefore does not handle such inputs as

THE LENGTH OF THE KENNEDY IS 1072 FEET  
LET THE LENGTH OF THE KENNEDY BE 1072 FEET  
SET THE LENGTH OF THE KENNEDY TO 1072 FEET.

Moreover, IDA itself does not provide for updating the database. Extending the language with new productions such as

<L.T.G> → SET THE <ATTRIBUTE> OF <SHIP> TO  
<VALUE>

would be easy, but there are serious database issues involved regarding consistency, security, and priority. Such database problems are beyond the scope of our research.

(e) *Irregular Coverage*. One of the consequences of the ease with which interface builders can add new patterns to a LIFER grammar is that gaps may appear in coverage. For example, suppose a given language definition contains no passive constructions. Through the use of paraphrase or by direct action on the part of the interface builder, the language may be extended to cover some, but perhaps not all, passive constructions. That is, the system might be made to accept

(1) THE KENNEDY IS OWNED BY WHOM,

but not

(2) THE KENNEDY IS COMMANDED BY WHOM.

(The semantically-oriented syntactic categories for OWNED and COMMANDED may differ.) If a user knows that the system accepts (1) and that the system accepts the active

**(3) WHO COMMANDS THE KENNEDY,**

then he is likely to be upset when input (2) is not accepted. In creating the language specification for INLAND, we have tried to minimize such irregularities in coverage by applying standard techniques of modular programming to the grammar specification. We feel this has been reasonably successful. Because LIFER gives the inference builder the freedom to add particular instance and subclasses of linguistic phenomena, it is his responsibility to avoid the gaps in coverage that may result.

*(2) Limitations Regarding Ambiguity.* The LIFER parser does not deal with syntactic ambiguity directly, but accepts its first successful analysis as being the sole interpretation of an input.<sup>10</sup> Because English contains truly ambiguous constructions, even when semantic considerations (the "augmentations") are taken into account, this limitation can be serious. For example, in the request

**(1) NAME THE SHIPS FROM AMERICAN HOME PORTS  
THAT ARE WITHIN 500 MILES OF NORFOLK**

the phrase THAT ARE WITHIN 500 MILES OF NORFOLK might modify either the SHIPS or the PORTS. The choice will, of course, influence the response made to the user. The current LIFER parser is biased against deep parses and will only consider the interpretation in which the clause modifies SHIPS. Even a single word can produce difficulties. For example, the word NORFOLK in request (1) could refer to a port in Virginia, a port in Great Britain, an American frigate, or a British destroyer. Thus the request is at least eight ways ambiguous. Codd (Codd 1974) has studied at some length the problem of ambiguity in the context of practical database systems and has developed the strategy of engaging in a dialogue in which the system articulates ambiguities (and other problems) and asks questions of the user to clarify the intent of his requests. In addition to the simple syntactic form of ambiguity, exemplified by request (1), other forms of ambiguity may arise. For example, the question

**(2) IS KENNEDY IN RADAR RANGE OF THE KNOX**

is syntactically unambiguous, but the meaning might be either

**IS KENNEDY IN KNOX'S RADAR RANGE**

or

**IS KNOX IN KENNEDY'S RADAR RANGE.**

---

<sup>10</sup>On October 31, 1977, LIFER was modified to allow optional production of all syntactically correct readings of an input. However, INLAND has not yet been revised to take advantage of this new option. When ambiguity is discovered, LIFER calls a user-defined subroutine with the list of parse trees (including response expressions and variable bindings at each non-terminal node) for all readings. One of the trees is to be returned for execution of the root level response expression. A default "user" subroutine is supplied with LIFER that prints the various parse trees and asks the user to select one by number. More sophisticated subroutines are expected to be written that will enter into more natural clarification dialogues.

This example represents a purely semantic ambiguity.  
Similarly,

### (3) IS KENNEDY NEARER TO GIBRALTAR THAN KITTY HAWK

might be considered syntactically unambiguous in its present form, yet it has two possible meanings. By adding "missing words" at two different points in the input it is possible to produce the readings

IS THE KENNEDY NEARER TO GIBRALTAR THAN  
the kennedy nearer to THE KITTY HAWK

and

IS THE KENNEDY NEARER TO GIBRALTAR THAN  
THE KITTY HAWK is near to Gibraltar.

Examples of ambiguity such as queries (1), (2), and (3) just given begin to show the difficulty of dealing with the problem in any general way. However, in the domain of INLAND, ambiguities have tended to arise only infrequently and have presented only minor problems for our particular application. Fortunately, our users have been very helpful by tending to avoid the use of the long and complex constructions that are most likely to lead to ambiguities. Perhaps this is because the teletype medium inclines users to prefer short, simple constructions. Even though LIFER does not deal with ambiguity directly, certain types of ambiguities may be trapped and treated by using the response expressions of LIFER production rules. For example,

```
PD[<L.T.G>
    (IS <SHIP1> IN <RANGE-TYPE> RANGE OF <SHIP2> )
    (COMPUTE.RANGE <SHIP1> <SHIP2> <RANGE-TYPE> )]
```

will accept such inputs as

IS KENNEDY WITHIN RADAR RANGE OF KNOX

and call the function COMPUTE.RANGE to respond.

COMPUTE.RANGE is given the two ships and the range type as an input. Knowing the pattern to be inherently ambiguous, COMPUTE.RANGE may enter into a (formal) conversation with the user to resolve the ambiguity.

The INLAND grammar also tries to avoid ambiguity whenever possible. For example, the phrase AMERICAN ARMORED TROOP CARRIER might mean a ship (of any nationality) that carries armored troops from the U.S. military, or an American ship that carries armored troops (of any nationality), or a ship that carries troops that were armored by the U.S., or a ship that is armored by the U.S. and that carries troops, or any one of several other combinations. In INLAND ARMORED-TROOP-CARRIER is recognized as a fixed phrase and all the problems with ambiguity vanish.

#### *(3) Limitations Regarding Definite Noun Phrases*

(a) *The Restricted Context Problem.* A phrase such as THE AMERICAN SUBS may be used to refer to different American submarines, depending upon the "con-

text" in which it appears. For example, if the Washington, the Churchill, and the Lincoln are being discussed, then THE AMERICAN SUBS in

## HOW OLD ARE THE AMERICAN SUBS

refers to the Washington and the Lincoln. Had the current context concerned the Roosevelt, Jefferson, and Leninsky Komsomol, then THE AMERICAN SUBS would have referred to the Roosevelt and the Jefferson. The point is that the meanings of certain noun phrases are dependent upon the contexts in which the phrases appear.

INLAND has only a limited ability to handle phrases such as THE SHIPS, THE AMERICAN SUB, and THOSE CRUISERS, which are said to be "definitely determined". As opposed to indefinitely determined noun phrases (such as A SHIP), which refer to the existence of objects not currently in context, definite noun phrases are often used to refer to a particular object or set of objects that is already in context. In dealing with a database, "in context" may usually be taken to mean "in the database". Thus the phrase THE AMERICAN SUBS generally means "the American subs in the database", and this is the interpretation that INLAND almost always places on this phrase. But suppose the user has just asked WHAT SUBS ARE IN THE MEDITERRANEAN and has been answered by a list of several subs, some of which are American and some of which belong to other countries. If the user now asks WHAT ARE THE POSITIONS OF THE AMERICAN SUBS he expects only the positions of American subs in the Mediterranean, but is given information about all American subs in the database. The problem is that the local context established by previous questions is more restricted than the total database and INLAND has not received enough lexical and syntactic clues to recognize this. (Had the input been WHAT ARE THE POSITIONS OF THE AMERICAN ONES, the use of the pronoun would have signaled the local context and INLAND would have replied properly.)

Where the context is very clear, INLAND can sometimes handle a restricted perspective on the database. For example, following

SELECT A MAP OF THE NORTH ATLANTIC

the query

DISPLAY THE AMERICAN SUBS

will cause the retrieval of only those subs in the North Atlantic, because others could not be displayed on the map in any case.

We know of no applied language system that deals adequately with this problem. However, significant experimental results are described in Grosz (Grosz 1977).

(b) *Some Methods for Treating Pronouns.* Even though the general problem of properly resolving pronouns is quite difficult, simple techniques can cover a large number of cases. For example, there are many trivial uses of pronouns in which no resolution is needed at all. Examples include

## WHAT TIME IS IT 1968 WHEN THE KENNEDY WAS LAUNCHED

and instances in which the pronoun references an earlier phrase in a pattern as in

WHEN WILL <SHIP> <HAVE> ITS <PART> <REPAIRED>  
(e.g. WHEN WILL THE KENNEDY GET ITS RADAR FIXED).

Very often pronouns are used in natural language queries to refer to things mentioned in the previous question. Thus in the sequence

WHAT IS THE LENGTH OF THE KENNEDY WHAT IS HER SPEED,

the pronoun HER refers to THE KENNEDY. Suppose the first sentence above is interpreted by means of the production

<L.T.G> → <PRESENT> THE <ATTRIBUTE> OF <SHIP>

and the second sentence by

<L.T.G> → <PRESENT> <SHIP'S> <ATTRIBUTE>.

The primary method for matching <SHIP'S> might be through a production such as

<SHIP'S> → <SHIP> <-'S>

where <-'S> is the possessive-forming suffix which is stripped off by a preprocessor.<sup>11</sup>

This primary method may be extended so that <SHIP'S> may also match HER or ITS or THEIR if a <SHIP> was used in the last input. This will allow WHAT IS HER SPEED to be interpreted as WHAT IS KENNEDY'S SPEED.

To extend the definition of <SHIP'S> to match pronouns, first define a predicate SHIP.PRONOUN that will return a non-NIL value if its argument is a possessive pronoun and the last input contained a <SHIP>. The predicate may be defined as

(LAMBDA (WORD)

    (AND (MEMBER WORD'(HER ITS THEIR))  
        (LIFER.BINDING'<SHIP> )))

where LIFER.BINDING is a LIFER function that determines whether the metasymbol given as its argument had a binding in the interpretation of the last input and, if so, returns the binding.<sup>12</sup> Using predicate SHIP.PRONOUN, the definition of metasymbol <SHIP'S> may be extended by the call

---

<sup>11</sup> Alternatively, a set of possessive nouns naming ships could be defined and the stripper not used, or, as is the case in INLAND, possessives could be thrown away altogether and <SHIP'S> could be made equivalent to <SHIP>.

<sup>12</sup>If there were multiple occurrences of the symbol in the last input, the leftmost-topmost instance is returned.

MP(<SHIP'S> SHIP.PRONOUN).

Another technique, which works nicely for some classes of anaphoric references, involves the use of global variables (sometimes called "registers"). For example suppose that each response expression associated with a pattern defining the metasymbol <SHIP> is so constructed that it will set the global variable LATEST-SHIP to the value it returns as the binding of <SHIP>. To be concrete,

```
PD[<SHIP>
  (<SHIP-NAME> )
  (SETQ LATEST-SHIP
    (LIST (LIST 'NAM 'EQ<SHIP-NAME> )))]
```

causes <SHIP> to match a <SHIP-NAME> as defined previously. The response expression that computes the value of <SHIP> will return the same value as defined above, but, as a side effect, it will now also set the global variable LATEST-SHIP to the same value. Later, when phrases such as THE SHIP or THAT SHIP are used to refer to the last ship mentioned, the global variable LATEST-SHIP may be used to recall that ship. For example, if <DET-DEF> is defined to match definite determiners (e.g. THAT, THE), then

```
PD[<SHIP>
  (<DET-DEF> SHIP)
  LATEST-SHIP]
```

will define structures that allow <SHIP> to match THE SHIP and take as its value the value of the LATEST-SHIP. Note that LATEST-SHIP is always ready with the value of the latest <SHIP> mentioned, but (LIFER.BINDING'<SHIP> ) is of help only if <SHIP> was used in the last input.

(4) *Limitations in Processing Elliptical Inputs.* After successfully processing the complete sentence

(1) HOW MANY CRUISERS ARE THERE

LIFER will accept the elliptical input

(2) CRUISERS WITHIN 600 MILES OF THE KNOX

but not

(3) WITHIN 600 MILES OF THE KNOX.

The elliptical processor is based on syntactic analogies. Input (2) is a noun phrase which is analogous to the noun phrase CRUISERS of input (1). Input (3), on the other hand, is a modifier that is intended to modify the CRUISERS of input (1). Because input (1) has no modifiers, elliptical input (3) has no parallel in the original input and hence cannot be accepted.

(5) *Other Limitations.* A few other important limitations of INLAND and LIFER are worth mentioning briefly. First, LIFER has no "core grammar" that is ready to

be used on any arbitrary database. This is because LIFER was designed as a general purpose language processing system and makes no commitment whatever to the types of programs and data structures for which it is to provide a front end or even to which natural language is to be accepted. LIFER might, for example, be used to build a Japanese language interface to a program that controls a robot arm. This could not be done if assumptions had been made restricting LIFER to database applications and to the English language. Thus LIFER contrasts with systems such as Thompson and Thompson's REL (rapidly extendable language) (Thompson 1975), which provides a core grammar but which requires reformatting of data into the REL database.<sup>13</sup>

Some systems, such as ROBOT (Harris 1977a, Harris 1977b), use the information in the data base itself as an extension of the language processor's lexicon. The LIFER interface may do this also but need not. If one elects not to use the database as lexicon, and this choice was made in INLAND, then the lexicon must be extended whenever new values are added to the database that a user may want to mention in his queries.<sup>14</sup> The price of using the database itself as an extended lexicon is that the database must be queried during the parsing process. For very large databases, this operation will probably be prohibitively expensive.

INLAND, of course, is basically a question answerer that relies on a database as its major source of domain information. In particular, INLAND cannot read newspaper articles or other extended texts and record their meaning for subsequent querying. Moreover, although it is perfectly reasonable that the LIFER parser might be used for a text reading system, LIFER itself contains no particular facilities other than calls to response expressions for recording or reasoning about complex bodies of knowledge.

## 6.2 The Role of the Task Domain

The limitations presented in the last subsection would cause major difficulties in dealing with many areas of natural language application. However, for our particular application, the limitations did not prevent the creation of a robust and useful system. In the next few paragraphs, we briefly outline some of the key features of the application that simplified our task.

The creation of INLAND was greatly facilitated by the nature of the particular interface problem that was addressed—providing a decision maker with access to information he knows is in a database. Because the user is expected to know what kinds of information are available and is expected to follow the technical terms and styles of writing that are typical in his domain of decision making, we can establish strong predictions about a user's linguistic behavior and hence INLAND needs to cover only a relatively narrow subset of language.

A second factor in facilitating the creation of the natural language interface was the interface provided by the IDA and FAM components of the LADDER system. By providing a simplistic view of what is in fact a complex and highly intertwined collection of distributed data, IDA and FAM helped greatly in simplifying the LISP

---

<sup>13</sup>Fragments of foreign-language versions of INLAND have been used to access the naval database in Swedish and Japanese.

<sup>14</sup>Because LADDER accesses data over the ARPANET, we felt the overall system would be intolerably slow if the actual database was used at parse time.

response expressions associated with productions in the INLAND grammar.

In short, IDA allows the database to be queried by high-level information requests that take the form of an unordered list of two kinds of items: fields whose values are desired, and conditions on the values of associated fields. Using IDA, the INLAND grammar need never be concerned with any entities in the database other than fields and field values. Furthermore, because the input to IDA is unordered, the construction of segments of a call to IDA can be done while parsing lower-level metasymbols.

The performance of INLAND for a given user is also enhanced by the user's own, often subconscious, tendency to adapt to the system's limitations. Because INLAND can handle at least the most straightforward paraphrases of most requests for the values of any particular fields, even a new user has a good chance of having his questions successfully answered on the first or second attempt. It has been our experience that those who use the system with some regularity soon adapt the style of their questions to that accepted by the language specification. The performance of these users suggests that they train themselves to understand the grammar accepted by INLAND and to restrict their questions whenever possible to forms within the grammar. Formal investigation of this subjectively observed phenomenon might prove very interesting.

### 6.3 The Role of Human Engineering

Although the basic language processing abilities provided by LIFER are similar to those found in some other systems, LIFER embodies a number of human engineering features that greatly enhance its usability. These humanizing features include its ability to deal with incomplete inputs and to allow users to extend the linguistic coverage at run time. But, more importantly, LIFER provides easy-to-understand, highly interactive functions for specifying, extending, modifying, and debugging application languages. These features provide a highly supportive environment for the incremental development of sophisticated interfaces. Without these supporting features, a language definition rapidly becomes too complex to manage and is no longer extendable. With support, the relatively simple types of linguistic constructions accepted by LIFER may be used to produce far more sophisticated interfaces than was previously thought possible.

Creating a LIFER grammar that covers the language of a particular application may be thought of constructively as writing a program for a parser machine. All the precepts of good programming—top-down design, modular programming, and the like—are relevant to good design of a semantic grammar. A well programmed grammar is easy to augment, because new top-level patterns are likely to refer to lower-level metasymbols that have already been developed and shown to work reliably. Thus the task of adding new top-level productions to a grammar is analogous to the task of adding new capabilities to a more typical body of computer code (such as a statistics package) by defining new capabilities in terms of existing subroutines.

No matter how well programmed a grammar might be, as the complexity of the grammar increases, the interactions among components of the language specification will grow. This leads the language designer into the familiar programming cycle of program, test, and debug. With many systems for parsing and language definition, the cycle may take many minutes for each iteration. With LIFER, when a new production is interactively entered into the grammar, it is immediately usable for testing by parsing

sample inputs. The time required for the cycle of program, test, and debug is thus dependent on the thinking time of the designer, not the processing time of the system. Because the designer can make very effective use of his time, he can support, maintain, and extend a language specification of far greater complexity than would otherwise be possible. The basic parsing technology of LIFER is not really new. But the human engineering that LIFER provides for interface builders has allowed us to better manage the existing technology and to apply it on a relatively large scale.

## 6.4 Related Work

As indicated by the February 1977 issue of the SIGART Newsletter (Erman 1977) which contains a collection of 52 short overviews of various research efforts in the general area, interest in the development of natural language interfaces is widespread. Our own work is similar to that of several others.

The LIFER parser is based on a simplification of the ideas developed in the LUNAR parser of Woods and others (Woods 1970, Woods 1972). In particular, LIFER manipulates internal structures that reflect Woods's ATN formalism. Woods's parser was used as a component of a system that accessed a database in answering questions about the chemical analysis of lunar rocks. The system did not use semantically-oriented syntactic categories and the database was smaller and less complex than that used by INLAND although the database query language was more general than that accepted by IDA.

Woods's ATN formalism has been used in a variety of systems, including a speech understanding system [26], and the semantically-oriented systems of Waltz (Waltz 1975), Brown and Burton (Brown and Burton 1975), and Burton (Burton 76). These latter systems do not use the LUNAR Parser, but rather compile the ATN formalism into procedures that in turn perform the parsing operation directly, without using a parser/interpreter to interpret a mathematical formalism. Compilation results in greater parsing speed, which is of importance for many applications. However, compilation also makes personalization features, such as PARAPHRASE, much more difficult to implement and increases the time of the program-test-debug cycle.

The first natural language systems to make extensive use of semantic grammars were those of Brown and Burton (Brown 1975) and Burton (Burton 1976). These systems were designed for computer-assisted instruction rather than as interfaces to databases.

In work very similar to our own, Waltz (Waltz 1975) has devised a system called PLANES which answers questions about the maintenance and flight histories of airplanes. PLANES uses both an ATN and a semantic grammar. Apparently the system does not include a paraphrase facility similar to LIFER's. It does support the processing of elliptical inputs by a technique differing from our own and supports clarification dialogues with users.

The PLANES language definition makes less use of syntactic information than INLAND. In particular, PLANES looks through an input for constituent phrases matching certain semantically-oriented syntax categories. When one of these constituent phrases is found, its value is placed in a local register that is associated with the given category. Rather than attempt to combine these constituent into complete sentences by syntactic means, "concept case frames" are used. Essentially, PLANES uses case frames to decide what type of question has been asked by looking at the types and values of local registers

that were set by the input. For example, the three questions

WHO OWNS THE KENNEDY  
BY WHOM IS KENNEDY OWNED  
THE KENNEDY IS OWNED BY WHOM

would all set, say, an <ACT> register to OWN and a <SHIP> register to KENNEDY. The case frames can determine what question is asked simply by looking, at the registers. Performing a complete syntactic analysis such as INLAND does require different constructions for each question pattern.<sup>15</sup>

If the input following one of the three questions asked in the preceding paragraph is the elliptical fragment "KNOX," the <SHIP> register is reset. Because no case frame is associated with <SHIP> alone and because <SHIP> was used in the last input, the <ACT> register is inherited in the new context and the elliptical input properly analyzed. When more than one case frame matches an input, PLANES enters into a clarification dialogue with the user to decide which was intended. (This conversation prints interpretations of inputs in a formal query language.)

The use of case frames is very attractive in that it allows many top-level syntactic patterns to be accounted for by a single rule. However, it is inadequate for complex inputs. The question IS KNOX FASTER THAN KENNEDY contains two <SHIP>s. Only the syntax tells us which to test as the faster of the two. Compound-complex sentences would be extremely difficult to process without extensive use of syntactic data. Waltz is investigating ways of supplementing his case frames with nominal pieces of syntactic information.

Codd's concept of the RENDEZVOUS system (Codd 1974) for interface to relational databases provides many ideas concerning clarification dialogues that might be included in LIFER at some later date. RENDEZVOUS is failsafe in that it can fall back on multiple choice selection if natural language processing fails completely.

Another applied natural language system whose underlying philosophy is akin to that of LIFER is the REL system developed by Thompson and Thompson (Thompson and Thompson 1975). REL is a data retrieval system like LADDER, though REL requires data to be stored in a special REL database. The grammar rules of REL contain a context-free part and an augmentation very much like those of LIFER. As its name implies, REL was intended to be easily extendable by interface builders. Much effort has gone into making REL run rapidly and it is almost certainly faster than LIFER. However, this speed was gained by a low-level language implementation with the unfortunate side effect that response expressions are not easily written.

Recently, the Artificial Intelligence Corporation introduced a commercial product called ROBOT for interfacing to databases. As described in Harris (Harris 1977a), ROBOT "calls for mapping English language questions into a language of database

---

<sup>15</sup>LIFER may be used to support case frames, although this was not done in INLAND particular, <L.T.G> may be defined as an arbitrary sequence of <CONSTITUENT>s, where <CONSTITUENT> may be expanded as any of the semantically-oriented syntax categories used by the case system. The response expression associated with the expansions of <CONSTITUENT> cause global registers to be set, and the response expression associated with <L.T.G> →<CONSTITUENTS> may make use of these registers and the case frames in computing a top-level response. A case frame system supported by LIFER would, of course inherit LIFER's run-time personalization and introspection features.

semantics that is independent of the contents of the database". The database itself is used as an extension of the dictionary, and the structure of files within the database helps in guiding the parser in the resolution of ambiguities. Our own research indicates that the types of linguistic construction employed by users are rather dependent on the content of the database. We also worry that extensive recourse to a database of substantial size may greatly slow the parsing process, unless the file is indexed on every field. Moreover, our database is coded largely in terms of abbreviations that are unsuitable as lexical entries. Nevertheless, the notion of using the data itself to extend the capabilities of the language system is very attractive.

In addition to the work on near-term application systems, a number of workers are currently addressing longer-range problems of accessing databases through natural language. See, for example, Mylopoulos et al. (Mylopoulos et.al. 1977), Sowa (Sowa 1976), Walker et al. (Walker et. al. 1977), and Sacerdoti (Sacerdoti 1977). There are, of course, many people engaged in research in the general area of natural language processing, but a survey of their work is beyond the scope of this paper.

## 7 Conclusion

We have described a system called LADDER that provides natural language access to a large, distributed database. We have shown that the language processing component of this system, although based on simple principles and subject to certain limitations, is sufficiently robust to be useful in practical applications. Moreover, we have indicated that LADDER is not an isolated system but that other applied language systems have achieved significant levels of performance as well, particularly in interfacing to databases. We believe that the evidence presented indicates clearly that, for certain restricted applications, natural language access to databases has become a practical and practicable reality.

## 8 Glossary

DBMS	Database management system.
FAM	File access manager. Maps generic file names onto specific file names on specific computers at specific sites. Initiates network connections, opens files, and monitors for certain errors.
IDA	Intelligent data access. Presents a structure-free view of a distributed database.
INLAND	Informal natural language access to Navy data. The natural language interface to IDA, which incorporates a special-purpose LIFER grammar.
LADDER	Language access to distributed data with error recovery. Our total system composed of INLAND, IDA, and FAM.
LIFER	Language interface facility with ellipsis and recursion. The general facility for creating and maintaining linguistic interfaces.

MP	Make predicate. The LIFER function for defining a metasymbol as a set of lexical items.
MS	Make set. The LIFER function for defining a metasymbol as a set of lexical items.
VLDB	Very large database.

## 9 Appendix. An example session with LADDER

Please type in your name: TOD S.

Do you want instructions? (type FIRST LETTER of response) No

Do you want to specify a current location (default = Norfolk)? No

Do you wish distance/direction calculations to default to  
GREAT CIRCLE, or RHUMB LINE? (you can override by specifying in the query)  
Great Circle

1.What is the current position of the Kennedy?

PARSED!

Parse time: .68 seconds

\* This counts cpu time used by INLAND.

IDA: ((? PTP) (? PTD) (NAM EQ 'KENNEDY% JF'))

\* This is the call to IDA.

Connecting to Datacomputer at CCA1:

\* FAM indicates which computer is being accessed. The next

\* 13 lines are interactions between FAM and the Datacomputer.

>> ;0031 771108184236 IONETI: CONNECTED TO SRI-KL-22700010

>> ;J150 771108184238 FCRUN: V='DC-4/10.00.1' J=3 DT='TUESDAY, NOVEMBER  
8, 1977 13:42:38-EST' S='CCA'

>> 10041 771108184239 DNCTNX: DATACOMPUTER GOING DOWN IN 905 MIN BE-  
CAUSE SYSTEM IS GOING DOWN AT WED NOV 9 77 5:00:00AM-EST FOR 240 MIN  
DUE TO SCHEDULED PM

>> ;J200 771108184239 RHRUN: READY FOR REQUEST

#> Set parameters

#< Exit

CCA1:Z

#> Set parameters

#< V Verbosity (-1 to 5): 1

#< P PROCEED with Datalanguage [confirm with <CR>]

# The connection has now been established. FAM now logs in

# and opens the necessary files.

CCA1:LOGIN %TOP.ACCAT.GUEST :

CCA1:OPEN STOP.ACCAT.SAGALOWICZ.NSTDPORT1 WRITE;

CCA1:OPEN STOP.ACCAT.NTRACKHIST READ;

CCA1:OPEN STOP.ACCAT.NNSHIP READ;

```

# FAM now transmits the query.
CCA1:FOR R1 IN NNSHIP WITH (NAM EQ 'KENNEDY JF') FOR NSTDPORT1 , R2 IN
CCA1:NTRACKHIST WITH R2.UICVCN EQ R1.UICVCN BEGIN
STRING1 = R2.PTP STRING2 = R2.PTD END;
#> Total bytes transferred: 27
IDA = ((PTP '6000N03000W' PTD 7601171200))
    # This is the value returned by IDA.
Computation time for query: 4.077 seconds
    # This counts cpu time used by IDA and FAM. Extra time is
    # needed to establish the network connection, log in, and
    # open files.
Real time for query: 224.725 seconds
    # This measures real time from the time the request is made
    # to IDA until IDA returns the answer.
(POSITION 6000N03000W DATE 7601171200)
    Kennedy was last reported to be at 60 degrees North,
    # 30 degrees West, at noon on January 17, 1976.

```

2\_of kitty hawk  
Trying Ellipsis: WHAT IS THE CURRENT POSITION OF KITTY HAWK  
Parse time: .97 seconds

```

IDA: ((? PTP) (? PTD) (NAM EQ 'KITTY% HAWK'))
CCA1:FOR R1 IN NNSHIP WITH (NAM EQ 'KITTY HAWK') FOR NSTDPORT1 , R2 IN
CCA1:NTRACKHIST WITH R2.UICVCN EQ R1.UICVCN BEGIN
STRING1 = R2.PTP STRING2 = R2.PTD END;

#> Total bytes transferred: 27
IDA = ((PTP '3700N01700E' PTD 7601171200))
Computation time for query: 1.077 seconds
Real time for query: 78.105 seconds
(POSITION 3700N01700E DATE 7601171200)

```

3\_To what country does each merchant ship in the north atlantic belong  
PARSED!  
Parse time: .386 seconds

```

IDA: ((? NAT) (? NAM) ((TYPE EQ 'BULK') OR (TYPE EQ 'TNKR'))
(PTPNS EQ 'N') (PTPEW EQ 'W') ((PTPY GT 600) OR (PTPX LT 3600) OR
(PTPI GT 3900)) (? PTP) (? PTD))
CCA:OPEN STOP ACCAT SAGALOWICZ.NSTDPORT2 WRITE;
CCA:OPEN STOP ACCAT NNMOVES READ;
CCA1:FOR R1 IN NNMOVES WITH ((TYPE EQ 'BULK') OR (TYPE EQ 'TNKR')) FOR R2
IN
CCA1:NTRACKHIST WITH (PTPNS EQ 'N') AND (PTPEW EQ 'W') AND
CCA1: ((PTPY GT 600) OR (PTPX LT 3600) OR (PTPX GT 3900)) AND R2.UICVCN EQ

```

CCAI:R1.UICVCN FOR NSTDPRT2 , R3 IN NNSHIP WITH R3.UICVCN EQ R2.UICVCN  
BEGIN  
CCA1:STRING1 = R2.PTP STRING2 = R2.PTD STRING3 = R3.NAT STRING4 = R3.NAM  
END;

#> Total bytes transferred: 4582

IDA = ((NAT 'US' NAM 'GREENVILLE VICTORY' PTP '1533N03010W' PTD  
7601171200) (NAT 'US' NAM 'TOULE% J' PTP '1533N03010W' PTD 7601171200)  
...)

Computation time for query: 5.702 seconds

Real time for query: 228.231 seconds

I have 79 answers; do you want to see them all? ...Yes

NAT SHIP POSITION DATE

US GREENVILLE VICTORY 1533N03010W 7601171200

US TOULE J 1533N03010W 7601171200

. {other entries omitted here}

WG TRANSQUEBEC 4633N02129W 7601171200

NO TALABOT 0600N01130W 7601171200

4\_the fastest sub

Trying Ellipsis: TO WHAT COUNTRY DOES THE FASTEST SUB BELONG

Parse time: 1.484 seconds

IDA: ((? NAT) (? NAM) (\* MAX MCSF) (TYPE1 EQ 'S') (TYPE2 EQ 'S'))

\* In the interests of brevity and clarity, the remaining

\* interactions with the Datacomputer are omitted.

IDA = ((NAT 'US' NAM 'LOS% ANGELES' MCSF '30.0'))

Computation time for query: 2.317 seconds

Real time for query: 173.3 seconds

(NAT US SHIP LOS ANGELES MXSPD 30.0)

5\_what ships are en route to america

Trying Ellipsis: Ellipsis has failed.

The parser does not expect the word "AMERICA" to follow

"WHAT SHIP ARE EN ROUTE TO"

Options for next word or symbol are:

<COUNTRY>, <PORT>

end-of-list

# LIFER cannot parse the question; "America" is not in

# the vocabulary.

6.what is a <country>

PARSED!

# In addition to questions about the data base, INLAND can  
# answer questions about the grammar itself.

"<COUNTRY>" may be any sequence of words following one of the patterns:

<COUNTRY> => THE <COUNTRY>

<COUNTRY> => U S

<COUNTRY> => U S S R

<COUNTRY> => U S S

<COUNTRY> => U S A

<COUNTRY> => U K

<COUNTRY> => SOVIET UNION

<COUNTRY> => UNITED STATES

<COUNTRY> => UNITED KINGDOM

<COUNTRY> => SOUTH AFRICA

<COUNTRY> => WEST GERMANY

<COUNTRY> => SAUDI ARABIA

<COUNTRY> => GREAT BRITAIN

<COUNTRY> => H M S

"<COUNTRY>" may be any member of the set {ANGOLA ANGOLAN ARABIA ARABIAN ARGENTINA ARGENTINAN BRITAIN BRITISH CANADA CANADIAN DUTCH EGYPT EGYPTIAN ENGLAND ENGLISH FOREIGN FRANCE FRENCH GERMAN GERMANY H.M.S. HMS ITALIAN ITALY LIBERIA LIBERIAN NETHERLANDS NORWAY NORWEGIAN PORTUGAL PORTUGUESE RUSSIA RUSSIAN SOVIET SPAIN SPANISH U.K. U.S. U.S.A. U.S.S. U.S.S.R. UK US USA USS USSR VENEZUELA VENEZUELAN)

Finished

7.define america like usa

PARSED!

# The user may add new synonyms to the vocabulary.

FINISHED

8.redo 5

# Here we are using the "redo" feature of INTERLISP.

PARSED!

Parse time: .356 seconds

IDA: ((? NAM) (DSC EQ 'US'))

IDA = ((NAM 'KENNEDYS JF') (NAM 'LOS% ANGELES') (NAM 'BATON% ROUGE') (NAM 'PHILADELPHIA') (NAM 'POGY') (NAM 'ASPRO') (NAM 'SUNFISH') (NAM 'KAWISHIWI'))

Computation time for query: 1.098 seconds

Real time for query: 67.16 seconds

SHIP = KENNEDY JF, LOS ANGELES, BATON ROUGE, PHILADELPHIA, POGY, ASPRO, SUNFISH, KAWISHIWI

9.\_how many of them are navy ships

THEM => ((DSC EQ 'US'))

PARSED!

# 'Them' or 'she' is currently always interpreted as a  
# reference to a set of ships in the previous query.

Parse time: .505 seconds

IDA: ((? NAM) (DSC EQ 'US') ((TYPE NE 'BULK') AND (TYPE NE 'TNKR'))))

IDA = ((NAM 'KENNEDY% JF') (NAM 'LOS% ANGELES') (NAM 'BATON% ROUGE')  
(NAM 'PHILADELPHIA') (NAM 'POGY') (NAM 'ASPRO') (NAM 'SUNFISH') (NAM  
'KAWISHIWI'))

Computation time for query: 1.205 seconds

Real time for query: 89.417 seconds

8 Or them:

SHIP = KENNEDY JF, LOS ANGELES, BATON ROUGE, PHILADELPHIA, POGY, ASPRO,  
SUNFISH, KAWISHIWI

10.\_give status kitty hawk

Trying Ellipsis: Ellipsis has failed.

The parser does not expect the word "STATUS" to follow

"GIVE"

Options for next word or symbol are:

<RELATIVE.CLAUSE>, <SHIP>, <VALUE.SPEC>, THE  
end-of-list

11.\_define (give status kitty hawk)

like (list the employment schedule, state of readiness, commanding  
officer and position of kitty hawk)

PARSED!

# This is an example of the paraphrase feature of LIFER. A  
# new pattern is defined by example.

Parse time: .705 seconds

# The system answers the query as a side-effect of parsing  
# the paraphrase.

IDA: ((? ETERM) (? EBEG) (? EEND) (? READY) (? RANK) (? CONAM) (? PTP) (? PTD)  
(NAM EQ 'KITTY% HAWK'))

IDA = ((ETERM 'SURVOPS' EBEG 760103 EEND 760205 READY 2 RANK 'CAPT'  
CONAM 'SPRUANCE% R' PTP '3700N01700E' PTD 7601171200))

Computation time for query: 2.725 seconds

Real time for query: 173.404 seconds

(EMPLMT SURVOPS EMPBEG 760103 EMPEND 760205 READY 2 RANK CAPT NAME  
SPRUANCE R POSITION 3700N01700E DATE 7601171200)

LIFER.TOP.GRAMMAR => GIVE STATUS <SHIP>

```
# The generalized pattern for the paraphrase is added to  
# the grammar  
F0086 (GIVE STATUS <SHIP>)  
# F0086 19 the new LISP function created to be the response  
# expression for this pattern.
```

12\_give status us cruisers in the mediteranean

spelling-> MEDITERRANEAN

PARSED!

Parse time: 2.855 seconds

IDA: ((? ETERM) (? EBEG) (? EEND) (? READY) (? RANK) (? CONAM) (? PTP) (? PTD)  
(? NAM) (NAT EQ 'US') (TYPE1 EQ 'C') (TYPE2 NE 'V') (TYPE NE 'CGO'))

IDA = ((ETERM 'CARESC' EBEG 760101 EEND 760601 READY 1 RANK 'CAPT' CONAM  
'MORRIS% R' PTP '4000N00600E' PTD 7601171200 NAM 'CALIFORNIA')  
(ETERM 'CARESC' EBEG 751231 EEND 760615 READY 1 RANK 'CAPT' CONAM 'HARMS%  
J' PTP '3700N01700E' PTD 7601171200 NAM 'DANIELS% Jnnn') ...)

Computation time for query: 3.738 seconds

Real time for query: 195.698 seconds

EMPLMNT: CARESC CARESC CARESC CARESC

EMPBEG: 760101 751231 751231 751231

EMPEND: 760601 760615 760615 760615

READY: 1 1 1

RANK: CAPT CAPT CAPT CAPT

NAME: MORRIS R HARMS J EVANS O FRENZINGER T

POSITION: 4000N00600E 3700N01700E 3700N01700E 3700N01700E

DATE: 7601171200 7601171200 7601171200 7601171200

SHIP: CALIFORNIA DANIELS J WAINWRIGHT JOUETT

{information about 8 other ships omitted}

13.done

PARSED!

# The user indicates that he is finished with the session.

File closed 8-Nov-77 11:11:17

Thank you

@

### Acknowledgments

Our debugging of LIFER and the continuing development of human engineering features have been strongly influenced by interactions with interface builders. In particular, we would like to thank the following people for using the LIFER system extensively and sharing their experiences with us: Staffan Lof (Swedish version of LADDER and extensions to INLAND), Martin Epstein (medical database system of melanoma cases), and Harry Barrow and Keith Lantz (interactive aid for cartography and photo interpre-

tation). We would also like to thank Gordon Novak for recent revisions to the elliptical processor.

## References

- [1] Brown, J.S., and Burton, R.R., "Multiple representations of knowledge for tutorial reasoning", in D.G. Bobrow and A. Collins, (eds.), *Representation and Understanding*, Academic Press, New York, 1975, 311-349.
- [2] Burton, R.R., "Semantic grammar: An engineering technique for constructing natural language understanding systems", BBN Rep. 3453, Bolt, Beranek, and Newman, Boston MA, Dec. 1976.
- [3] Codd, E.F., "Seven steps to rendezvous with the casual user", in J.W. Klimbie and K.I. Koffeman, (eds.), *Data Base Management*, North-Holland, Amsterdam, 1974, 179-200.
- [4] Computer Corporation of America. Datacomputer version 1 user manual. CCA, Cambridge MA, Aug. 1975.
- [5] Erman, L.D., (ed.), *ACM SIGART Newsletter*, 61, Feb. 1977.
- [6] Farrell, J., "The Datacomputer-A network data utility", in *Proc. Berkeley Workshop on Distributed Data Management and Computer Networks*, Berkeley, CA, May 1976, 352-364.
- [7] Grosz, B.J., "The representation and use of focus in dialog understanding." Ph.D. dissertation, U. of California, Berkeley, May 1977.
- [8] Harris, L.R., "ROBOT" A high performance natural language processor for data base query." *ACM SIGART Newsletter*, 61, Feb. 1977, 39-40.
- [9] Harris, L.R. "User oriented data base query with the ROBOT natural language query system," *Proc. 3rd Int. Conf. on Very Large Data Bases*, Tokyo, Japan, Oct. 6-8, 1977.
- [10] Hendrix, G.G., "The LIFER manual: A guide to building practical natural language interfaces." Tech. Note 138, SRI Artificial Intelligence Center, Menlo Park, CA, Feb. 1977.
- [11] Hendrix, G.G., "Human engineering for applied natural language processing", *Proc. 5th Int. Joint Conf. on Artificial Intelligence*, Cambridge MA, August 1977.
- [12] Hopcroft, J.E., and Ullman, J.D., *Formal Languages and Their Relation to Automata*, Addison-Wesley, Reading MA, 1969.
- [13] Morris, P. and Sagalowicz, D., "Managing network access to a distributed data base", *Proc. 2nd Berkeley Workshop on Distributed Data Management and Computer Networks*, Berkeley CA, May 1977.

- [14] Mylopoulos, J., Bordgida, A., Cohen, P., Roussopoulos, N., Tsotsos, J. and Wong, H., "TORUS-A natural language understanding system for data management", *Proc. 4th Int. Joint Conf. on Artificial Intelligence*, Tbilisi, U.S.S.R., Aug. 1977.
- [15] Paxton, W.H., "A framework for speech understanding", Tech. Note 142, SRI Artificial Intelligence Center, Menlo Park CA, June 1977.
- [16] Sacerdoti, E.D., "Language access to distributed data with error recovery", *Proc. 5th Int. Joint Conf. on Artificial Intelligence*, Cambridge MA, Aug. 1977.
- [17] Sagalowicz, D., IDA: An intelligent data access program," *Proc. 3rd Int. Conf. on Very Large Data Bases*, Tokyo, Japan, Oct. 1977.
- [18] Sowa, J.F., "Conceptual graphs for a database interface," *IBM J. Res. Develop.*, 20 4, July 1976, 336-357.
- [19] Teitelman, W., INTERLISP reference manual. Xerox PARC, Palo Alto CA, Dec. 1975.
- [20] Thompson, F.B., and Thompson, B.H., "Practical natural language processing: The REL system as prototype," in M. Rubinoff and M.C. Yovits, (eds.), *Advances in Computers 13*, Academic Press, New York, 1975.
- [21] Walker, D.E., Grosz, B.J., Hendrix, G.G., Paxton, W.H., Robinson, A.E., and Slocum, J., "An overview of speech understanding research at SRI," *Proc. 5th Int. Joint Conf. on Artificial Intelligence*, Cambridge MA, Aug. 1977.
- [22] Waltz, D., "Natural language access to a large database: An engineering approach," *Proc. 4th Int. Joint Conf. on Artificial Intelligence*, Tbilisi, U.S.S.R., Sept. 1975, pp. 868-872.
- [23] Woods, W.A., "Transition network grammars for natural language analysis," *Comm. ACM 13*, 10, Oct. 1970, 591-606.
- [24] Woods, W.A., Kaplan, R.M., and Nash-Webber, B., "The lunar sciences natural language information system," BBN Rep. 2378, Bolt, Beranek and Newman, Cambridge MA, 1972.
- [25] Woods, W.A., "An experimental parsing system for transition network grammars," in R. Rustin, (ed.), *Natural Language Processing*, Algorithmics Press, New York, 1973.
- [26] Woods, W.A., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., Makhoul, J., Nash-Webber, B., Schwartz, R., Wolf, J., and Zue, V., "Speech understanding systems," Final technical progress report, Tech. Rep. 3438, Bolt, Beranek, and Newman, Cambridge MA, Dec. 1976.

# User-Needs Analysis and Design Methodology for an Automated Document Generator\*

K. Kukich †  
K. McKeown ‡  
J. Shaw ‡  
J. Robin ‡  
J. Lim ‡  
N. Morgan †  
J. Phillips †

†Bellcore  
445 South Street  
Morristown, NJ 07960-6438  
*e-mail:* [kukich@bellcore.com](mailto:kukich@bellcore.com)

‡Columbia University  
Computer Science Department  
New York, NY 10027  
*email:* [mckeown@cs.columbia.edu](mailto:mckeown@cs.columbia.edu)

## Abstract

Telephone network planning engineers routinely study feeder routes within the telephone network in order to create and refine network capacity expansion (relief) plans. In doing so they make use of a powerful software tool called *LEIS™ – PLAN*<sup>1</sup>. We are developing an extension to PLAN called PLANDoc that will automatically generate natural language narratives documenting the engineers' use of PLAN. In this paper, we present the user-needs analysis and design methodology we have used in developing the PLANDoc system. We describe our interviews with various end users to determine if such a system would be desirable and what design factors would make it useful. We show how we model the system on a set of iteratively-revised human-generated narratives. The model narratives determine the function and architecture of the documentation system, and they inform the development of the system components.

---

\*Appeared in: Proceedings of the Fourth Bellcore/BCC Symposium on User-Centered Design, November 3-5, 1993, Piscataway, NJ.

<sup>1</sup>LEIS is a registered trademark of Bellcore, Livingston, NJ.

# 1 Introduction

User-needs analysis has become a common practice in the development of computer-human interface systems and other end-user software (Nielsen 1993), (Landauer et al. 1993). However, AI systems are frequently developed without an analysis of the end user's needs. Particularly in developing a large scale, practical AI system, the needs of the user should be studied if the resulting system is to be effectively used. In this paper, we report on a user-needs analysis and system development methodology that we are using in our ongoing development of an automated documentation system, PLANDoc, for telephone network planning operations. We are systematically modeling PLANDoc after a target set of manually-written narratives that have undergone successive revisions based on user critiques. We show how user input has influenced both the form and content of PLANDoc narratives and the design of the system itself.

The telephone network planning engineer's job is to derive a capacity expansion (relief) plan specifying when, where, and how much new copper, fiber, multiplexing and other equipment to install in the local network to avoid facilities exhaustion. Planning engineers have the benefit of a powerful software tool, the Bellcore *LEIS<sup>TM</sup> – PLAN*, that helps them derive a 20-year relief plan that optimizes the timing, placement and cost of new facilities for a route in the network. Until now they have not had the benefit of a tool to help them with the equally important but often tedious task of documenting their planning decisions, so in many busy shops this crucial writing task has been upstaged by more pressing planning tasks. Documentation is needed primarily to provide a record of the planner's activities and reasoning to be used for future network studies, for informing managers who are responsible for authorizing project plans, and for justifying expenditures to internal auditors and external regulators.

Our work enhances PLAN by providing a natural language text generation system and a simple facility that prompts engineers to add their own manually-written comments at crucial points. Based on the results of our user-needs analysis, PLANDoc is designed to combine human-generated and machine-generated text to produce natural language narratives. In the following sections we briefly describe the planning engineer's use of the PLAN system, our user-needs analysis and the resulting PLANDoc narrative format, the PLANDoc system architecture and the development of the text generator based on target narratives. We close with a discussion of future research challenges related to content planning for PLANDoc.

## 2 LEIS-PLAN Background

Voice and data service is carried to telephone customers through a complex network of routes consisting of copper or fiber cables supplemented by additional equipment such as Digital Loop Carrier (DLC) systems and fiber multiplexors. In order to be able to provide new services to customers on demand, network facilities must already be in place. It is the planning engineer's job to determine the optimum configuration of equipment needed to avoid facility exhaustion while minimizing costs.

The engineer uses PLAN to download a model of the route from a headquarters database, to enter forecasts of anticipated service demands, and then to have PLAN compute an optimum, cost-effective base relief plan. The base plan specifies the timing,

PLANDoc Tracking Report	PLANDoc Refinement Paragraph with Engineering Note
Run-ID reg1 for DLC PLAN Changes [Activ] CSA 3122 used ALL-DLC system idlc272 and DSS-DLC system idlc272 CSA 3122 for ALL-DLC activated 3 quarter 1994 for DSS-DLC activated 3 quarter 1994 CSA 3130 for ALL-DLC activated 3 quarter 1994 for DSS-DLC activated 3 quarter 1994 CSA 3134 for ALL-DLC activated 3 quarter 1994 for DSS-DLC activated 3 quarter 1994 CSA 3208 for ALL-DLC activated 3 quarter 1994a for DSS-DLC activated 3 quarter 1994 CSA 3420 for ALL-DLC activated 3 quarter 1994 for DSS-DLC activated 3 quarter 1994 Total 20 Year PWE (\$K) For The Route: \$2110 First 5 Year IFC (\$K) For The Route: \$1064	Run-ID Reg1: This refinement activated CSA's 3122, 3130, 3134, 3208 and 3420 for DLC in the third quarter of 1994. DLC system idlc272 was used for all placements in CSA 3122. [The route default system is 'idlc96'.] For this refinement, the resulting 20-year route PWE was \$2110K, a \$198K savings over the base plan, and the 5-year IFC was \$1064K, \$65K penalty over the base plan.  Engr Note: These CSA's are beyond 28 kf and need range extenders to provide service on copper. Moving them to 1994 will negate a job adding a reg bay to the office.

Figure 1: Tracking report example

placement and costs of new facilities needed to meet forecast demand over the next twenty years, with particular emphasis on the next five years.

Although PLAN employs sophisticated optimization algorithms to arrive at the most economical long-term base plan, the base plan may not always be realizable or desirable. It frequently needs to be refined to account for political, economical, practical and other factors known to the engineer but unknown to the computer. For example, the planning engineer may need to make changes to accommodate a management policy such as a mandate to achieve an all-fiber route by a certain date; she may need to work within the constraint of a fixed cap on initial installation cost; or she may be aware of practicalities such as the existence of small manholes that make equipment installation difficult in certain locations.

For this reason the PLAN system includes a powerful Interactive Refinement Module that allows the engineer to do 'what-if' modeling to explore the effects of various changes to the base plan. Some of the refinement actions an engineer might explore include requesting copper or fiber cable placements, requesting a DLC activation for a given site, changing a fiber activation time, or requesting a fiber service extension, among others. After comparing the effects of different refinement scenarios in terms of costs, timing and placements of equipment, the engineer ultimately decides on a realizable relief plan to recommend to management for project authorization.

PLAN keeps a somewhat cryptic trace of all the refinements the engineer tries in a refinement Tracking Report. Figure 1 shows a portion of a Tracking Report for one refinement together with its corresponding refinement paragraph generated by PLANDoc and its manually-entered engineering note.

### 3 User-Needs Analysis and Model PLANDoc Narratives

With the help of Bellcore Planning and Engineering staff<sup>2</sup> we formulated an initial proposal for PLANDoc and drafted preliminary target narratives. We then conducted a series of interviews with planning engineers and managers and PLAN support staff from several regional telephone companies in their home offices and at two PLAN training courses<sup>3</sup>. The work experience of the engineers we interviewed ranged from beginner to expert. Our goal was to determine 1) how engineers actually used the PLAN system, 2) whether they would find an automated documentation facility to be helpful, and, if so, 3) what the form and content of the narratives should be, and 4) how the documentation system should function. Finally, we compiled a corpus of target narratives written by an experienced planning engineer, incorporating feedback from the interviews.

Some of the things we learned about how engineers use PLAN include: a) novice planners often run exploratory refinements just to develop a feel for the effects of making certain changes in the route; b) experienced planners sometimes run refinements they know will be suboptimal just for the record, i.e., for the benefit of managers, auditors and regulators who might ask "did you try such and such?". More critical to the need for documentation, we also learned that: c) experienced planners keep handwritten notes on paper listing their refinements and why they tried them; they asked for a way to enter their notes on-line to keep track of their reasoning; d) inexperienced planners tend to rely on the base plan because they lack the expertise to refine it; they asked to see narratives written by experienced planners in order to learn from them; (unfortunately few such narratives exist); finally, e) all planners welcomed the idea of having the computer generate narratives that they could include in their documentation packages, especially if they could add to the narratives themselves.

These findings shaped the content of PLANDoc narratives and the design of the system. Specifically, they indicated that planners may not want all the refinements they tried, to appear in the narrative. For example, novice planners do not want to include their exploratory refinements, while experienced planners do want to include the suboptimal refinements they ran to show that their final refinements were superior. Thus, PLANDoc includes a facility that lets the planner select a subset of refinements to be included in the final narrative. Planners made it clear that they use knowledge not included in PLAN to make their decisions (e.g., knowledge of other installation projects, corporate strategies, physical environment, etc.) and they wanted a way to record that knowledge on-line, while they were working. This gave rise to PLANDoc's facility to prompt for manually-written engineer's notes at crucial points. Throughout the PLANDoc design process, we took care to minimize and standardize changes to PLAN's original, successful interface. Thus, PLANDoc requires only two modifications to PLAN's interface, one to prompt for engineering notes and another to allow the engineer to request a narrative and select a subset of refinements to be included. Both options are presented using familiar PLAN interface commands and screen formats.

---

<sup>2</sup>Many thanks to M. Horwath, D. Imhoff and L. Tener

<sup>3</sup>Some of the helpful regional Planning and Engineering personnel included P. McNeill, J. Bruner, P. King, D. Kelly, I. McNeill, T. Smith, C. Lowe, and G. Giles from Pacific Bell, R. Riggs, D. Spiegel, S. Sweat, L. Doane, R. Tufts, and R. Ott from Southwestern Bell, S. Wasalinko from NYNEX, and C. Lazette from Ameritech.



Figure 2: PLANDoc system architecture

In addition to our interviews, we also arranged for an experienced retired planning engineer, Jim Phillips, who is also a PLAN expert, to write a corpus of target narratives based on PLAN runs of actual routes. We have been using this corpus to guide the development of the PLANDoc text generator. Based on the findings from our interviews and on successive revisions and critiques of target narratives, we arrived at the following general model for PLANDoc narratives that integrates machine-generated text with the engineer's manually-written notes:

### **Model PLANDoc Narrative Format**

PART 1: Tabular Route Input Data Summary

PART 2: Base Plan Summary Paragraph

PART 3: Alternating Paragraphs of Refinement Descriptions and Engineer's Notes

PART 4: Relief Plan Summary Paragraph

## **4 The PLANDoc Automatic Text Generator**

PLANDoc's architecture, which is shown in Figure 2 draws on our previous report generation and text generation work (Kukich 1983), (McKeown 1985).

Since most of PLANDoc is written in LISP, a message generator<sup>4</sup> produces a LISP representation for each refinement action in a PLAN tracking report. This set of messages is first passed to an 'ontologizer'<sup>5</sup> that enriches each message with semantic knowledge from PLAN's domain. The messages are then passed to a content planner<sup>6</sup> whose job it is to determine which information in the messages should appear in the various summaries and to organize the overall narrative. For our initial prototype, we have focused only on the generation of paragraphs that describe refinement actions (Model Narrative, PART 3), and not on the generation of the summary paragraphs. Thus, currently the content planner's task is limited to combining individual messages to produce the input for complex sentences. We are using an existing surface generation system, the FUF/SURGE package (Elhadad 1991), (Elhadad 1993) as a tool in the text generator. It consists of a unifier (FUF) and large grammar of English (SURGE) that enables it to generate a large variety of sentences. In order to generate the sentences required to "translate" the refinement actions, we built a lexical interface<sup>7</sup> which maps

---

<sup>4</sup>written by N. Morgan

<sup>5</sup>written by J. Robin and J. Shaw

<sup>6</sup>written by J. Robin and J. Shaw

<sup>7</sup>written by J. Shaw with input from J. Robin, D. Radev, J. Lim, M. Elhadad and D. Horowitz

the tokens in the messages to the input required by FUF/SURGE. Since FUF/SURGE accepts a case role structure as input, with words already chosen, the lexical interface must determine the semantic roles and words in the sentence for each token in the input message.

Our user needs analysis directly influenced our system development. For example, an analysis of PLAN's menu of refinement actions and the model narratives allowed us to specify the different possible message types. Furthermore, a systematic categorization of the sentences in our model narratives revealed all the different phrasings for each message type. This categorization showed that there is great variety in the possible sentences for each message type. For example, all message types that specify a type of refinement that the planning engineer carried out could be realized as "This refinement demanded <action np>", as in "This refinement demanded DLC activation for CSA 2119 in 1994 Q1." Alternatively, these sentences could be realized as "<Site X> <passive action verb>" as in "CSA 2119 was activated for DLC in 1994 Q1.". These are but two of many possibilities. The model narratives indicate that which of these alternative phrasings is used depends on whether any refinements for this site have already been mentioned in the summary (i.e., the choice is based on previous discourse). If none had been mentioned, the summary used "This refinement demanded ..." and if the site was just discussed, the summary used the verb for the action in the passive form.

Thus, the use of the model narratives for system development has provided us with the data for lexicon development, both in terms of vocabulary and sentence structure used. In addition, it provides us with the constraints on alternative realizations. By examining the narratives, we can find the situations in which different forms and words were used. Currently, the implemented text generator can produce sentences for 24 out of a total of 32 different message types. It can produce an average of 150 different sentences for a single message and the lexicon contains 145 open class words in addition to equipment database terms and the closed class words which are selected by the grammar. It is able to produce complex sentences which combine several messages as shown in the output of Figure 1. The first sentence in this text combines 5 activations of DLC using conjunction.

## 5 Conclusion and Current Directions

The user-needs analysis we described here has influenced many aspects of system development. The model narrative format has determined the functions of the PLANDoc system by specifying the types of documentation required. Interviews with the planning engineers has influenced the design of PLANDoc's interface, resulting in facilities that allow engineers to select which refinements of the ones they tried should be included in the report and to add in their own notes which will be integrated with the automatically generated text. Finally, the manually-written target narratives have guided the development of the PLANDoc content-planner and lexicon. They have provided us with the data which allowed us to systematically identify the different sentences forms and words that are needed to describe each refinement action as well as to identify constraints that determine when different sentence types should be used.

Although the design of a content planner and a lexical grammar for managing the constraints of discourse structure lexical choice has posed many interesting research

challenges, other difficult challenges remain to be addressed in the design of an even more intelligent narrative generator. These challenges are primarily related to content planning. One challenge is to enrich refinement paragraphs with relevant information drawn from PLAN's input data, base plan and previous refinements. An example of this is shown in the bracketed sentence included in Figure 1. Another challenge is to generate summary paragraphs for the base plan and relief plan. This challenge requires developing models for determining which few of the many facts included in the data are relevant to the readers' needs (where readers include engineers, managers, auditors and regulators.) As with the design of the current content planning and lexicalization modules, we intend to exploit the corpus of target narratives, supplemented by interviews with expert planners, to derive the models needed for generating more intelligent refinement paragraphs and summary paragraphs.

### Acknowledgments

As this article goes to press, the alpha version of the PLANDoc system is being tested (many thanks to K. Slator, J. Phillips, W. Conkright, R. Parker) in preparation for release to a trial site in one of the Regional Bell Operating Companies (many thanks to M. McShane). Reaching even this preliminary stage of technology transfer would have been impossible without the support and contributions of the large number of research, development, administrative and subject matter experts mentioned throughout this article. In addition, we are especially grateful to Michael Lesk, Lynn Streeter and John Kaminski for throwing their managerial support behind us. But our greatest debt is to Don Walker, whose incredible energy first brought us together and inspired us to make this practical text generation application a reality. His legacy will fire the imagination and influence the practice of computational natural language processing for ages to come.

## References

- [1] Elhadad, M., "FUF: The universal unifier user manual version 5.0", Tech. Report CUCS-038-91, Computer Science Department, Columbia University, New York, 1991.
- [2] Elhadad, M., "Using argumentation to control lexical choice: a unification-based implementation", PhD. Thesis, Computer Science Department, Columbia University, New York, 1993.
- [3] Kukich, K., "The design of a knowledge-based report generation", in *Proceedings of the 21st Conference of the ACL*, University of Pittsburgh, 1983.
- [4] Landauer, T. et al., "Enhancing the Usability of Text Through Computer Delivery and Formative Evaluation: The SuperBook Project", in Dillon, A. and McKnight, C. (eds.), *Hypertext: A Psychological Perspective*, 1993.
- [5] McKeown, K., *Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*, Cambridge University Press, 1985.
- [6] Nielsen, J., *Usability Engineering*, Academic Press, San Diego, CA, 1993.

**SECTION 2**

**BUILDING COMPUTATIONAL LEXICONS**

# Machine-Readable Dictionaries and Computational Linguistics Research\*

Branimir Boguraev  
Apple Computer, Inc.  
Advanced Technologies Group  
*e-mail: bkb@apple.com*

## Abstract

A substantial amount of work has gone into attempts to provide computational solutions to a wide range of problems which rely on information extracted from machine-readable dictionaries (MRDs). Syntactic parsing, grammar development, word sense selection, speech synthesis, robust text interpretation, knowledge acquisition, information management, phonetically guided lexical access, phrasal analysis — these are only a few of the language processing functions which have been substantially aided by the availability of large dictionaries on-line. The paper looks at representative examples from this list and analyses them within a framework which aims to establish broad categories within computational linguistics to which MRDs have been (or could be) applied, and modes of their use. In addition to questions to do with the nature of the coupling between the contents of a dictionary entry and the task to which this is applied, I will use the same case studies to address a class of issues at a different level of generality — these include the reliability of an MRD from an application's point of view, the cost of extracting the information relevant to a computer program, and the optimal structure and organisation of a machine-readable source.

## 1 Background

This paper has three points of departure, and before getting into further details, I will attempt to establish the premises for the discussion which follows, the boundaries

\*This paper was written, upon Don Walker's request, for an international workshop on "Automating the Lexicon: Theory and Practice", which took place in 1986 in Grosseto, Italy. The workshop was a culmination of several years of effort, primarily and largely on Don's part, to bring together researchers from many disciplines concerned with a range of matters and issues lexical, and to define a broad platform for lexical research. It can hardly be said that this is where the lexicon was recognised as an important component in the larger framework of computational linguistics/natural language processing. It is certainly true, however, that it was at this meeting that, among other things, the foundation was laid to what today is generically referred to as *computational lexicology* and, more specifically, *automated lexicon acquisition* from machine-readable sources. The original paper is reprinted here, with a grateful acknowledgement to Oxford University Press, in a somewhat revised and edited form. However, the changes in the field since the Grosseto meeting have been so profound, that incorporating them in the body of the paper would have been impossible without substantial modification. Instead, a brief perspective reflecting the developments since 1986 is available in the appendix of this paper.

within which it will be contained, and the dimensions along which some presentations, comparisons and evaluations will be made.

## 1.1 Why Machine-readable Dictionaries?

A great number of systems within the computational linguistics paradigm are strikingly limited in their range of application. This lack of versatility is to a large extent due to the typically small number of lexical entries available to them. A moderately recent workshop on linguistic theory and computer applications (Whitelock et al., 1987) reports on an informal poll to establish the average size of the lexicon used by the prototypes discussed. Taking into account a five to six thousand word vocabulary used by a machine translation system, the average lexicon size came to 1500 words. Without this vocabulary, the average size was about 25. A related observation concerns the fragility of most experimental systems in the face of serious attempts to ‘scale up’ laboratory prototypes or application programs carefully tuned for particular domains. Some of the problems here are connected to issues of lexicon acquisition, knowledge base elicitation, robustness and so forth, and I shall be looking at these in more detail below.

Admittedly, most of these are not production type systems; nonetheless, the fact remains that for real world applications, especially given the recent advances in computational linguistics technology which make such applications feasible, considerably larger vocabularies are required. Given a number of constraints, ranging across a wide variety of systems — the most critical of which are idiosyncratic formalisms for writing dictionary entries, different lexicon formats, and diverse views on what constitutes linguistically relevant (syntactic, semantic, and pragmatic) information whose proper place is in the lexicon — the task of generating a comprehensive and consistent vocabulary for any application looks, and indeed is, quite awesome.

It is hardly surprising, then, that a number of researchers have been looking at available machine-readable dictionaries (MRDs). They expect that information already digested, categorised, indexed and, most importantly, available in machine-readable form, can be suitably used, if not to get a sizable lexicon ‘for free’, then at least to construct automatically a substantial portion of such a lexicon, getting for free what is hoped to be an internally consistent (and coherent) object. A lexicon derived in such a way would save considerable effort and still produce the bulk of the target vocabulary which could be subsequently expanded, if necessary, and further tailored to the task and application at hand.

This paper will investigate some ways in which these goals are being pursued by a variety of researchers, using different dictionaries and utilising different techniques to support different applications. The emphasis will be, as the title suggests, on analysing how information available in machine-readable form has been, and could be, used in computational linguistics research.

## 1.2 What is Computational Linguistics?

In attempting to define the scope of the term *computational linguistics* (CL) as it applies to the main area of concern of the paper, I shall follow the outline of the discipline, as presented by Thompson (1983), where computational linguistics is seen to occupy an

area of convergence between linguistics, psychology, philosophy and computer science. Thompson further offers the following breakdown: computation in service to linguistics, computational psycholinguistics, computation in service to psychology, theory of linguistic computation, computational linguistics proper and applied computational linguistics. It is only the last two classes that will provide the context for this investigation; thus, for example, I shall not attempt to cover such areas as the use of machine-readable dictionaries in the setting of psycholinguistic experiments or in theoretical/computational lexicography. On the other hand, insofar as applied computational linguistics might offer convenient tools and techniques to neighbouring disciplines, e.g., information retrieval or knowledge representation, these will also figure in the discussion.

However, I will not be looking at specific well-defined tasks within CL, e.g., parsing, generation or speech recognition. (Generation at this moment of time seems to be a relatively understudied subject anyway, and not enough is known about the knowledge and processes governing it to warrant text production of sufficiently large proportions, itself stimulating work on use of MRDs for generation. An exception here might be the JANUS project at USC/ISI with its Nigel generation component — see Cumming, 1986a.) Instead, the relevance of a dictionary entry will be examined with respect to certain (local) problems related to these tasks, either within the scope of computational linguistics, as defined above (e.g., lexical disambiguation or morphological analysis) or requiring CL techniques (for example knowledge-based spelling correction).

### 1.3 The Nature of a Dictionary Entry

How does the general notion of a dictionary entry relate to a generic entry in a lexicon for a natural language processing system? What kinds of information is it reasonable to expect to find in a machine-readable dictionary, which might be utilised profitably by some application program? While it is true that systems differ in the organisation, structure and content of their lexicons, it is still possible to isolate certain types of information, to be determined by the particular task and application, which ought to be made available at the lexical level. A categorisation of the lexical requirements, as they vary across a range of applications within the framework of computational linguistics, will provide a dimension for evaluating the utility of lexical information that may be available from a machine-readable source.

At a certain, uninteresting, level of abstraction, most dictionaries offer at least the following types of information: phonetic, syntactic and semantic. It is the further decomposition of these broad classes, however, that is of more interest here. Below is a convenient classification of the information in a dictionary entry used as a starting point for this enquiry.

The *form* contains headword information, spelling, syllabification and phonetic description, with possible variants in the spelling or pronunciation. Additional associated information might be to do with glosses on use of the word (e.g., formal, slang, etc.), variations in its spelling upon inflection, its behaviour when undergoing derivational morphology changes, stress patterns and so forth.

The entry's *function* describes its distributional behaviour. Here we might have a simple word class tagging (for example the *Collins English Dictionary*, Hanks, 1979, only lists broad grammatical classes), or a very elaborate grammatical subcategorisation

of the item in the style of the *Oxford Advanced Learner's Dictionary of Current English* (Hornby, 1974, henceforth OALDCE) or *Longman Dictionary of Contemporary English* (Procter, 1978, henceforth LDOCE).

The *meaning* would be given by means of one or more definitions, examples, and cross references. Some dictionaries might supply even more information about the usage of the word, both grammatically and stylistically, perhaps punctuated by comparisons with synonyms, antonyms, related words, and so forth. Occasionally, pictorial or diagrammatic information might be provided, etymologies supplied, and various other devices used to offer further insights into the word's meaning and patterns of use. Not directly related to the meaning, but within the same broad class, would be pointers to derived words, compound terms, idiomatic or common phrases and expressions naturally indexed under the headword and collocations.

In general, the distinction drawn here between *form*, *function* and *meaning* is not as clear cut as it appears. Properties of a word, e.g., its syntactic subcategorisation patterns, are very often a guide to its meaning(s). The categorisation presented above is intended to facilitate subsequent analysis of the different classes of use to which some of the information in a dictionary entry has been applied.

It is unlikely that most applications making use of a machine-readable source will require access to all of this information. It is also unlikely that until further progress is made in the more general aspects of natural language understanding, significant use can be made on-line of, say, cross reference pointers, comparisons between antonyms and synonyms, or pictorial information<sup>1</sup>. Machine-readable sources are still only reflections of what are, essentially, objects intended primarily for human consumption. Printed dictionaries can afford to be quite informal, as it suits them, being able to rely to an enormous extent on their readers' judgement and intelligence. We are a long way away from being able to automatically formalise facts like "There is no noun formed from **fast** when it means **quick**. Use instead **speed** or **quickness**".

On the other hand, most of the phonetic, distributional and definitional information to be found in a dictionary entry can be quite useful to a range of application programs, and it is these uses that I shall examine in more detail below. However, where the program in question is, in fact, a natural language processing system, it is a sad fact that every system has its own ideas and conventions concerning the content, organisation, and structure of its lexicon. Such a state of affairs is partly justified by differences, organisational or theoretic, in the individual systems' approaches to, e.g., parsing or generation (see Ingria, in this volume, and Cumming, in this volume). Still, this makes it impossible to share linguistically relevant information across systems. Until a rich, powerful and flexible lexical database is developed, where different linguistic theories can find relevant lexical information as they need it, building customised lexicons will inevitably involve duplication of effort. In the meantime, where general purpose lexicons will be developed and released for wide use, these will have deliberately uncommitted, and certainly uninteresting, lexical entry structure: for example, a project which has produced a dictionary and morphological analysis system for English to be used in conjunction with a parser for Generalised Phrase Structure Grammar (Gazdar et al.,

---

<sup>1</sup>There is some work, however, e.g., at Bell Communications Research (Bellcore), studying the relationship between a dictionary entry and its associated information; research is also being carried out investigating the application of videodisc technology for the display of dictionary illustrations under computer control.

1985, henceforth GPSG) defines in its user guide the following structure for an entry: "each entry is a 4-tuple — citation form, syntactic category, semantic *name*, and a *user field*, and is intended to represent a single morpheme" (Ritchie et al., 1987, emphasis added).

The question of how a dictionary entry relates to its lexical counterpart in a natural language processing system has a 'flip side'. Computational linguistics research does not always follow linguistic canons, and occasionally offers models for computer based language processing which not only do not relate in a direct (or obvious) way to any particular linguistic theory, but pose specific requirements on processing configurations and lexicon content and structure.

Small's theory of word expert parsing (Small, 1980), for instance, places a heavy demand on the lexicon; the lexical entries (word experts) there are complicated programs with coroutine structure, the specification of which requires detailed knowledge of the architecture of the parser, acquaintance with a specialised language for writing word experts, judgement of what constitutes linguistically relevant information and how to represent that procedurally, and readiness to bring in arbitrary amounts of more general, common world, knowledge. The point here is not so much whether the theory in question has any value, but that this is an instance of a lexicon for a natural language processing system which could not be derived from any machine-readable dictionary in existence. Similarly, recent tendencies to exploit parallel architectures result in various schemes for connectionist parsing (see, for example, Cottrell and Small, 1983 and Waltz and Pollack, 1985), whose implementations would require a qualitatively different type of lexicon.

Sowa and May (1986) present a different theory of text interpretation in context, relying on the notion of a conceptual graph. The implications for the content and structure of a lexicon to be used by a conceptual parser are non-trivial, and below I shall be looking at some general, and related, issues. Examples here include questions like: Where is the borderline between linguistic and real world knowledge with respect to text interpretation? How much of the shallow world knowledge to be found in a dictionary rather than in an encyclopaedia is required by such a parser? What would a taxonomy of concepts and a specification of the relationships between them look like? It is far from clear how easily an MRD can be tailored to suit such a theory; on the other hand, work is being done on using a machine-readable source to provide practical answers to some of these questions.

A different point to note here concerns one type of information, particularly relevant to virtually any kind of linguistic processing, and yet not to be found in a dictionary. Dictionary users are assumed to possess knowledge about morphology, both inflectional and derivational. While a reasonable assumption to make, given the reality of human linguistic performance and the practical issues of keeping printed dictionaries within reasonable size, this makes it a requirement for any client computer application to be able to perform at least some sort of morphological processing. An interface to a machine-readable source must regard this source as a dynamic object and must provide, in addition to the static data encoded as a set of morphemes with associated features held in the on-line dictionary, morphological processes together with enabling software for real-time word analysis perhaps carried over a local area network (see, for example, Kay, 1984 and Domenig, 1986). It was the desire to be able to utilise more fully the information in an on-line version of the *American Heritage Dictionary*, that motivated

the work of Kay and Kaplan (1981) on finite state morphology; this, in its own right led to Koskenniemi's computationally tractable model of two-level morphological analysis (1983), itself the basis for considerable efforts to implement linguistically motivated and computationally efficient morphological analysers (see, for example, Russell et al., 1986).

## 2 MRD-Based Research

The purpose of this paper is not to present an exhaustive survey of all computational linguistics research which is based on using a machine-readable dictionary. This would be well beyond the scope of this presentation, as well as impossible to do because of in-house research of a more proprietary nature carried out in some industrial environments. The emphasis here is going to be on examining and categorising the kinds of language processing functions and the tasks they can be applied to, and by extension, the type of research made possible by the existence of dictionaries in machine-readable form. No strong preference is made for a particular dictionary, and while certain names are mentioned more often than others, this only indicates closer familiarity with the corresponding machine-readable sources.

Instead of enumerating individual projects, I will attempt to present in broad outline the different types of activity based on utilising the information in an MRD. There are at least two more or less orthogonal dimensions for consideration in such an evaluative survey.

On the one hand, I am concerned with the nature of information extracted from a dictionary entry. The distinction sought here is the distinction brought out in the previous section, namely the one between *properties* and *meanings* of words. In the former case, a machine-readable source can be viewed as a lexical base, offering information about, for instance, the syllabification of a word or its idiosyncratic syntactic behaviour. Such information might be relevant to a speech synthesis system or a parsing program, but is of very little utility to, say, the interpretation component of a natural language interface. In the latter case, the intention is to regard the dictionary as a sort of a knowledge base, where information of more general semantic nature would be encoded, either in a free text format, or in some form of semantic tagging. Of general interest in both cases are issues like the kind of data that are likely to be available in machine-readable form, and the class of computational system for which such data may be useful.

The orthogonal question of how particular aspects of such data are extracted from the machine-readable source and made available to a client program defines the second dimension for this survey. In a typical case any raw information to be found in a dictionary is unlikely to be in a form that is immediately utilisable by a program, and connecting this to the original machine-readable source is not a trivial task. It is not only a matter of whether the dictionary is available on-line or not, with the immediate implications for extracting and restructuring selected fields of an entry 'on the fly'. Even assuming a proper access route to an entry, the finding, in the machine-readable source, of all the information required by the application program at any particular point of the processing still presents problems. What makes this a critical issue is the fact that depending both on the content of an entry and on the nature of the language processing function, the relevant information may or may not be extracted easily (or, indeed, at all)

from the dictionary — consider, for example, following a chain of word definitions in an attempt to place a particular concept in its proper place in a generalisation hierarchy. The distinction sought here is one between *first-* and *second-order utilisations* of an MRD where use is made, respectively, of *local* or *distributed* data. In the former case simply hooking a dictionary to a client program and carrying out suitable transformation(s) would suffice to support the task at hand. The latter would require processing the dictionary off-line and collating information from more than one place in it before deriving a data structure capable of being utilised by the client program.

Yet another, perhaps somewhat frivolous, dimension for categorising research based on available machine-readable sources is distinguishing between *problem-driven* and *interest-driven* work. The former category would include activities where commitment to a particular goal motivates the use of an MRD. The dictionary source here is viewed as a necessary tool and an indispensable component of the work. For example, a recent project in the UK (within the framework of the Alvey programme for information technology) contracted to deliver a computational grammar of English, developed within the framework of GPSG, together with a 50,000 strong word list indexed to the grammar (Boguraev et al., 1986). The project relied on having access to a machine-readable source of lexical data, suitably indexed and tagged (in the style of, e.g., LDOCE or the Lancaster-Oslo-Bergen corpus — see Garside and Leech, 1982 and Atwell et al., 1984), and would have been impossible to conceive otherwise. In contrast, a different strand of work in Cambridge (Alshawi, 1989) has been investigating the applicability of a particular text skimming technique to the task of dictionary definitions analysis, and a variety of ideas for subsequent use of the resulting system (see below) have been pursued in the last few months. Some of this work has been of an ‘opportunistic’ nature, motivated simply by the availability of a machine-readable source and a suite of programs capable of processing this in a certain way.

I will not attempt to follow this particular distinction too closely; it is worth bearing in mind, however, that research in computational linguistics which is predicated on MRDs of any sort falls in a different class from the more ‘pure’ kind of research, where work starts with fewer presuppositions, assumptions and constraints. No dictionary has been developed with a particular linguistic theory in mind. Consequently, it is inevitable that a certain amount of effort will be spent by each project in attempts to adapt the available information to its specific theory and framework; this is independent of whether the data are to be used to substantiate the theory or simply as a compendium of theory-free data. It is not clear if all such attempts are going to be successful, and this is where questions concerning ultimately the reliability and utility of MRDs for computational linguistics research become relevant.

### 3 The Use of MRDs in Computational Linguistics

Recent surveys of the state-of-the-art of natural language processing seem to agree on the class of possible application areas, regardless of whether the analysis is carried from the perspective of theoretical work within the artificial intelligence (AI) paradigm (see for example Waltz, 1983) or of purely commercial interests (Tim Johnson, 1985). A classification broad enough to cover the whole field includes machine translation, document understanding (or content scanning), document preparation aids, document

generation, systems control (where natural language capabilities are typically applied to interfacing to the system and maintaining a dialogue) and speech recognition and synthesis.

The utility, potential and real, of machine-readable dictionary sources to all of these areas is beyond question. In a survey of the field Amsler (1984) lists some of the tasks to which the on-line versions of the *Merriam-Webster Seventh New Collegiate Dictionary* (W7) and the *Merriam-Webster New Pocket Dictionary* (MPD) have been applied. These include student essay analysis and more general text understanding, stylistic analysis, parsing of dictionary definitions, experimentation in information retrieval, derivation of word lists to be used for spelling correction or hyphenation, and so forth. More recently, word lists for spelling correction have been derived by, e.g., Yannakoudakis (1983, 93,000 words extracted from the *Shorter Oxford Dictionary*) and Fawthrop and Yannakoudakis (1983, 57,000 words derived from the *Teachers Word Book of 30,000 Words* (Thorndike and Lorge, 1944)). Still within the general framework of Pollock and Zamora (1984), Mitton has initiated a project on using the phonetic information in the OALDCE to correct certain classes of misspelt words (Mitton, 1986, and personal communication). Word lists have also been generated from machine-readable sources to support machine translation (see for example Tucker and Nirenburg, 1984) and automated indexing (Klingbiel, 1985).

More recently MRDs have been applied to a range of related, as well as completely new, tasks. Cowie (1983), presents a system for analysing descriptive texts into hierarchically structured knowledge fragments; developments to that system make use of the machine-readable source of LDOCE. These include work on extracting a suitable lexicon for the system from the dictionary, as well as adapting it to handle textual fragments from the word definitions in it. Work with similar aims, namely the assembly of a 'folk taxonomy' of English plant and animal terms, has been carried out at Bell Communications Research (Walker and Amsler, personal communication); however, the strategy employed there is different from Cowie's: the definitions in W7 are not analyzed to any depth, but are scanned for various syntactic clues.

Other work at Bellcore has investigated methods for construction of, and browsing through, a network of sense relations, derived from dictionary definitions and, in particular, their synonym and cross reference pointers. This activity has been carried out in the context of George Miller's WORDNET project (Miller, 1985) whose goal is to develop a system, appropriately organised and indexed, and equipped with navigational aids for examining complex conceptual spaces without having to conform to the conventional alphabetic arrangement of dictionary data. The system is aimed at human users; a suitable counterpart, however, could be of great utility to a computer program for, e.g., robust interpretation of free text.

Exploring implicit connections between word senses and 'browsing' through the networks of concepts behind dictionary definitions has found use in lexical disambiguation. In contrast to procedurally based disambiguation techniques, where semantic processes are applied either concurrently with or after syntactic analysis of the input text (e.g., Boguraev, 1980; Hirst, 1987), methods for disambiguating multiple word senses in context have been developed which make use only of information found in a machine-readable source, and do not require any syntactic processing, thus lending themselves to more general application across language boundaries. This kind of approach can be

traced back to the work done in the Cambridge Language Research Unit (Masterman et al., 1957), where the problem of word sense disambiguation was tackled by exploiting the structure of *Roget's Thesaurus*. More recently, Amsler (1983) makes certain assumptions about the (usually implicit) taxonomic structure of the defining concepts in a dictionary; his approach relies on this structure being made explicit prior to the enterprise (not an unreasonable standpoint, as we shall see below). Lesk's method (1986) is even more direct, in that all it requires is a machine-readable source, and not of any particular dictionary at that. He uses heuristics depending on the overlap between words in the dictionary-provided sense definitions for the words across a context window.

The Lexical Systems group at IBM Yorktown Heights has been working on WordSmith, an automated dictionary system designed on top of W7 to offer similar browsing functionality: users can retrieve words which are 'close' to a given word along dimensions such as spelling, meaning and sound. Some of the tasks WordSmith has been applied to, namely developing techniques for segmenting and matching word spellings and generating pronunciations for unknown words, are described in Byrd and Chodorow (1985).

During the course of a speech synthesis project at Murray Hill, Church (1985) has also looked at the applicability of large machine-readable lexical sources (dictionaries and corpora) to the problem of stress assignment. Streeter (1978, and personal communication) and Coker have been explicitly tackling the problems associated with deriving a pronouncing dictionary from W7 for the purposes of speech synthesis. (Note that most of these are caused by the fact that dictionaries do not always give pronunciations for all forms of a word; consequently satisfactory solutions necessarily involve a good morphological analysis component, capable of handling derivational and inflectional morphology alike.) Briscoe (1985) lists various recent (and current) projects in speech recognition and synthesis in the UK (sited at Edinburgh University, Leicester Polytechnic, The Joint Speech Research Unit at Cheltenham, The University of Cambridge and the IBM Scientific Centre at Winchester), all of which make use of sources like OALDCE or Collins for the generic task of compiling special purpose word lists transcribed into project-specific phonemic alphabets, incorporating primary and secondary stress assignment and marking of syllable boundaries.

Recent work by Huttenlocher, Shipman and Zue of the Massachusetts Institute of Technology has investigated an alternative model of lexical access to the one commonly utilised by conventional approaches to speech recognition (Shipman and Zue, 1982; Huttenlocher and Zue, 1983). Instead of applying classical pattern-matching techniques, which become inadequate for tasks requiring large vocabularies, they have analysed the particular knowledge about language and speech available in dictionaries, and have developed a special representation of the speech signal based on sequential phonetic constraints. Classification of words along the phonetic categories defined there achieves the partitioning of a large lexicon (e.g., the 20,000 strong MPD) into relatively small equivalence classes, thus greatly reducing the space of possible word candidates on look-up and making the whole process of lexical access less sensitive to speaker variation and other variabilities in the speech signal.

Recent work in linguistics, and in particular developments in grammatical theory — for example Generalised Phrase Structure Grammar (GPSG) (Gazdar et al., 1985), Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982) — and on natural

language parsing frameworks — for example, Kay's Functional Unification Grammar (FUG) (Kay, 1984b), PATR-II (Shieber, 1985) — make it feasible to consider the implementation of efficient systems for the syntactic analysis of substantial fragments of natural language. Machine-readable sources have been found to be of substantial use in such contexts. Perhaps the most comprehensive application of an MRD for the task of parsing proper is the use of an on-line lexicon of well over 100,000 words, compiled from machine-readable sources (primarily W7), in the CRITIQUE (formerly EPISTLE; see Heidorn et al., 1982) project at IBM Yorktown Heights. The lexicon is used by a powerful parser which comprises part of a larger system for machine processing of natural language text in an office environment, and specifically geared to carry out grammar and style checking of business letters. The system makes use solely of syntactic information of a limited nature: entries are tagged with part-of-speech labels and only crudely subcategorised (e.g., transitivity of verbs would be indicated, but not a lot else).

The more recently developed grammar formalisms cited above (with the exception of CRITIQUE's parser which is based on Augmented Phrase Structure Grammar) generalise the notion of a non-terminal symbol from a simple atom to a complex-valued feature cluster. One of the implications for the lexicon in such systems is that there has been massive relocation into it of linguistic information which hitherto used to be carried by phrase structure rules. While not all of the MRDs carry such complex information in their entries, some sources provide elaborate syntactic tagging of the lexical items. In particular, OALDCE and LDOCE employ systems for grammatical coding (see Michiels, 1982, for a very comprehensive study of the LDOCE structure and grammar tagging practices; see also Akkerman et al., 1985, for precise analysis and comparison between the coding systems of LDOCE and OALDCE) which allow for detailed syntactic subcategorisation of individual word senses. It is only natural to attempt to use this information about the idiosyncratic distributional behaviour of words for syntactic analysis. Furthermore, Michiels (1982) presents an algorithm for deriving a generative grammar-like feature system which distinguishes between different semantic types of verb (raising, equi, and so forth). Recent work at the University of Cambridge (Alshawi et al., 1985) demonstrates the feasibility of developing a system capable of dynamic restructuring of the Longman grammatical codes into complex feature clusters for the use by a parser (within the PATR-II framework in this particular instance).

Since the system employed by LDOCE for grammatical tagging is based on the descriptive grammatical framework of Quirk et al. (1972), the mapping of the original dictionary format into a lexical entry appropriate to an LFG/GPSG/PATR parser is not trivial. This has motivated the recent efforts in Cambridge to extend and generalise the system (see Boguraev and Briscoe, 1987), paralleled by work at CSLI (Annie Zaenen, personal communication; see also CSLI Monthly, 1986, for a fuller description of CSLI's Lexical Project). The intention is to build a centralised lexicon, derived from a machine-readable source like LDOCE, and containing subcategorisation and other relevant information in a form not committed to any particular formal theory of grammar. This intermediate 'meta-lexicon' would then be piped through 'grammar macros' (*templates* in the PATR-II terminology) to generate a target lexicon for one's favourite grammar formalism. A somewhat similar goal is being pursued in the ASCOT project at the University of Amsterdam (Akkerman et al., 1985), which also seeks to translate the information in the grammar codes fields of LDOCE in a form subsequently

utilisable by a system for analysis and corpora-style tagging of text. Independent work which might be relevant to all of these projects is the research carried out by Jackendoff and Grimshaw (1985), which has produced a set of subcategorisation and control codes for English and has applied these to a non-trivial computer-accessible verb list.

The work in Cambridge is loosely related to the already mentioned ongoing project to deliver a computational grammar of English with a 50,000 strong word list indexed to it. The machine-readable source of LDOCE is going to be used in this project not only in the way indicated above, but also as an essential component of the overall grammar development environment. Beyond the extraction of a GPSG-orientated lexicon, starting from the Longman grammar codes and going through a series of (manually aided) iterations in the process of 'tuning' the resulting lexicon to the grammar, it is intended to use the examples associated with the word sense definitions for testing the grammar coverage.

A cursory survey of the state-of-the-art in language processing would probably conclude that MRDs have not been used in any application falling within the scope of using natural language as an interface medium for communicating with computational systems. In particular, no projects to build a natural language interface component of any kind have attempted to derive their lexicons from machine-readable sources. As Rich (1984) points out, there are essentially two reasons for this. Words 'mean' different things to practically oriented systems where the back-end applications have widely different sets of underlying concepts, relationships between them and commands to manipulate them. A general purpose dictionary is of very little utility when trying to establish precise mapping between, say, a natural language verb and its counterpart action (or property) in a relational database. Furthermore, words with no mapping in an application domain have no 'meaning' to the interface. Using an unabridged MRD source to derive the system's lexicon would then result in a large number of meaningless entries in the lexicon, and still would fail to provide a comprehensive coverage of the domain and task at hand. (N.B. However, see below for some work which might provide a solution to this particular problem.)

This is the reason why natural language interfaces have carefully hand-crafted lexicons where large amounts of idiosyncratic domain- and task-specific information is built in 'from scratch'. This is particularly true of front-end systems which have left the purely research environment to venture into the commercial world — the high performance required for their survival on the market can only be achieved by placing an enormous amount of information in their lexicons, which vary from installation to installation and could not, in any specific instance, be derived from an MRD.

This situation might change, under the recent work on transportable natural language interfaces (see Ballard and Biermann, 1985), and in particular given approaches where the front-end system makes extensive use of general purpose syntactic (e.g., the TEAM project: Martin et al., 1983) and semantic (e.g., Boguraev and Sparck Jones, 1983) analysers. The architecture of such systems incorporates substantial lexicons which can be transported across domains, together with the rest of the application-independent components. It would make sense to acquire most, if not all, of these lexicons from machine-readable sources. This is, in fact, the primary motivation behind the UK Natural Language Toolkit effort, already mentioned here (Boguraev et al., 1986; Ritchie et al., 1987): the resulting grammar, parser and lexicon are intended to be used by a

range of practical projects, most of them aimed at building customised natural language interfaces. Recent activity in the US (observed over the ARPANET and too unstructured to reference properly here) indicates that there is increased interest, both in the research and in the industrial worlds, in obtaining MRDs and extracting suitably sized lexicons from them.

There is still another aspect of the work on interfaces in general, to which research in making use of the information available from MRDs is particularly relevant. Latter day language processing programs fall within the larger category of 'knowledge based systems' in Artificial Intelligence. In order to carry out their (semi-) intelligent functions, such systems need significant amounts of structured knowledge about the real world, or at least about a particular domain of discourse. A common problem addressed during the design process is that of acquiring such knowledge, and there is strong hope that ways can be found to localise and extract some of it from suitable machine-readable sources, namely dictionaries or encyclopaedias. I shall address this issue in more detail in the next section.

Thus it would seem that the only application within the broad classification of natural language processing where MRDs have not found use so far is, as I already pointed out earlier, that of text generation (speech synthesis is excluded from this context)<sup>2</sup>. Just a quick glimpse at the lexical level problems to be solved for the process of generating text will suffice to demonstrate why MRDs have not been seriously exploited in this context. Ritchie (1987) mentions some of the questions to be considered during the lexical selection proper: what is the input to this process — a syntactic structure, a semantic representation, or a mixture of both? What are the dynamics of the lexical selection process, i.e. how does the preference for a particular word affect subsequent choice of words? What is the architecture of the lexical selection mechanism, i.e. at which point(s) of the transition from internal representation to surface text are references to the lexicon made?

Until motivated answers to these, and related, questions are found there will be no background for, and very little value in, attempts to bring MRDs in the process of language generation. Chances are that machine-readable sources, as we know them today, will not be able to supply all of the information required for language production, and a non-trivial amount of this information would have to be, at least initially, hand-crafted. I tend to agree with McDermott (1976) who suggests that no significant insights into natural language generation are going to be made until we have a program which has something significant to say — this seems ultimately to tie up the issue of use of MRDs for text generation with fundamental AI research into endowing intelligent programs with introspection capabilities.

## 4 The Information Content of a Dictionary Source

Most of the work mentioned in the previous section makes use of the form and function of a dictionary entry. The work on disambiguation and dictionary browsing appears to be

---

<sup>2</sup>Note that the process of language production from some underlying meaning representation should be distinguished from the functions carried out by various document generation aids, where MRDs have already been applied to on a large scale, e.g., for spelling correction within the CRITIQUE system.

an exception, in that the associated techniques and methods need access to a definition (or a subject code, synonym, or cross reference) field. This is, however, somewhat misleading, as the semantic *content* of the definition is not really made use of.

An analysis of the cases where heavy use has been made of the information in a dictionary entry ('heavy' in the sense that this information has been extracted *and* subsequently utilised by a client program of some sort), shows that the concern has typically been with the lexical properties of a word. Such information, normally to be found in the form and function fields, includes data on spelling, pronunciation, syllabification, perhaps (inflectional) morphology, grammatical distribution and subcategorisation. The MRD source in these cases is regarded as a *lexical base*, and the type of a client program that would need this lexical information usually knows quite precisely where to find it, how to extract it, and what, if any, transformations are required before making actual use of the raw data. (Note that the use of the term "lexical base" is quite different from what Amsler calls "lexical knowledge base" (Amsler, 1984b, and in this volume) as well as from Calzolari's "lexical data base" (Calzolari, 1984b).)

In contrast, work which tends to concentrate on making some use of the meaning component of a dictionary entry regards the MRD as a *knowledge base*; efforts here are directed at locating and extracting much more loosely defined, placed and structured raw data. Ultimately, the goal is to relate natural language words to an underlying taxonomy of concepts — typically the one which binds together the defining concepts in a dictionary. This would involve a range of activities, considerably more difficult than those which reduce to, e.g., simple lexical look-up, part-of-speech extraction, and its mapping into a data structure suitable for subsequent use by an on-line parser.

The problems under this heading are due to a variety of factors. Firstly, word meaning is typically given descriptively, by means of free text definitions. Apart from the fact that a parser for unrestricted natural language text still has not been developed, it has proved quite difficult to impose a formal structure on a wide set of dictionary definitions, as they have been written by different lexicographers, and are intended for human consumption, as opposed to rigorous analysis by a computer program.

Furthermore, the meaning of a single item (word sense) is not necessarily to be found only under the dictionary heading for this item. An arbitrarily long chain of pointers and cross references to other definitions may be involved. The further complications introduced by refining meaning by references to synonyms, antonyms and arbitrary associated information have already been mentioned, and since virtually no use is being made of them in computational linguistics research, I shall not discuss these here. A different problem, this time one which researchers must account for, is introduced by the possibility (and reality) of circularity in dictionary definitions. I shall address this issue in more detail later.

Finally, dictionaries are not encyclopaedias. Definitions, however detailed they may be, still assume some fundamental, general knowledge about the world. Attempts to extract the meaning of a word sense from its description in a dictionary and to convey this by means of an encoding in a formal knowledge structure require before anything else the formalisation of this general knowledge, without which no useful interpretation of any particular definition can be achieved. This poses the question of how to establish the common ground assumed by the lexicographers during the process of writing the definitions. (Note, however, that there is no implication that *all* the knowledge required

to understand a dictionary definition can be found in an encyclopaedia.)

Current views on automatic natural language processing tend to agree that there seems to be a continuum between the minimal semantic knowledge implied by the use of a particular word (word sense) and the specialised (or expert) knowledge relevant to its use in a given domain context (see, for example, Wilks, 1977; or more recently, Cater, 1987). For a practical natural language system, particularly of the kind discussed at the end of the last section, there are very pragmatic reasons for distinguishing between lexical semantics and specialised world knowledge. It would be unreasonable to expect to find any of the latter in a MRD. It would be of an enormous utility if most of the former — no matter whether it is presented in terms of decomposition into selectional restrictions markers (Katz and Fodor, 1963), formulae constructed from semantic primitives (Wilks, 1977), frame-based structures (Hirst, 1987), logical predicate/function symbols with associated sortal information encoded in the form of meaning postulates (Grosz and Stickel, 1983), or by some other means — could be derived from a machine-readable source.

In this context, there are attempts to compile some lexical semantics from a MRD into lexicons for a natural language processing systems. Wilks (personal communication) has recently initiated a project to use LDOCE for automatically constructing semantic definitions (formulae) for individual word senses, in the spirit of his approach to language analysis. This work is taking advantage of the fact that this particular dictionary offers constrained definitions, without circularities and using a limited defining vocabulary of not more than 2000 basic words. A similarly motivated project in Cambridge is using the word sense definitions and the selectional restrictions provided by the LDOCE tape to derive skeletal semantic formulae which can subsequently be fleshed out by hand and applied to domain-independent, semantically based language analysis.

The work, already mentioned, by Sowa and May (1986) is concerned with related issues. Some of the research there investigates a range of questions related to lexicon acquisition for a conceptual parser, where the lexical semantics takes on a very different form, and deriving conceptual graphs from a machine-readable source is by no means an easy problem.

Still, we should not expect to be able to generate a complete semantic component for a natural language processing system semi-automatically, in the same way in which we are attempting to flesh out the syntactic one. The semantic content of an MRD is going to be useful in a different fashion by offering access to a substantial body of facts potentially of relevance to AI applications.

Knowledge based programs organise and maintain their knowledge bases (KBs) in ways which differ widely across the range of their tasks and applications. A knowledge representation (KR) scheme suitable for a program that learns by analogy is not necessarily suitable for a backward chaining expert system. Even within the area of natural language processing there is no firm consensus on what kinds of structure are best suited for capturing the knowledge useful for language interpretation and understanding.

Nonetheless, it is possible to observe a common theme in a large number of language processing systems which, in addition to the relatively narrowly defined language-specific data, make use of more general, real-world, knowledge, as well as of more specialised, domain- and task-dependent knowledge. More often than not this kind of knowledge is represented using a scheme based on the general notions of frame-like

concepts with slot-like role descriptions, organised in an (inheritance) hierarchy along a generalisation/specialisation axis. Most of the recent work on knowledge representation, including FRL, KRL, NETL, AIMDS, UNITS, FRAIL or KL-ONE (Brachman and Schmolze, 1985), to mention but a few KR languages, can be cast into this general mould.

The utility, for natural language processing systems, of hierarchically structured networks of concepts, both relating to the real world and to specific domains, has been demonstrated beyond doubt (see, for example, Bobrow and Webber, 1980; Mark, 1981; Haas and Hendrix, 1983; Boguraev et al., 1986). It is not altogether clear whether all of the structured knowledge required for the operation of such systems can be derived in a systematic and consistent way from a machine-readable source. A substantial project at MCC (Lenat et al., 1985) is engaged in developing (manually) a large knowledge base of real world facts and heuristics, extracted from an encyclopaedia. At the Imperial Cancer Research Fund in London, John Fox (personal communication) is collaborating with the Oxford University Press to use the *Oxford Dictionary of Medicine* for the derivation of a medical KB suitable for use by a range of expert systems. It is yet to be seen how successful these, and related, projects will be.

On the other hand, dictionaries seem to offer a very convenient starting point for an initial compilation of taxonomically structured knowledge. Following some work of more empirical nature in the sixties, Amsler (1980), in the course of a comprehensive investigation into the content and structure of MPD, presented conclusive evidence that a dictionary contains a non-trivial amount of information which can semi-automatically be structured in a semantic hierarchy of defining concepts. This hierarchy could then be utilised by a number of computer programs — his own technique for lexical disambiguation, for example, relies on having access to such a taxonomy of 'genus' terms.

Since then, Calzolari (1984a), Chodorow et al. (1985) and Alshawi et al. (1985, 1986), among others, have done more work on analysing dictionary definitions with view of extracting a variety of relationships among lexical entries and concepts behind them. Significantly, one common thread in these projects is the effort to automate the locating of superordinates (genus extraction) without having to invoke a full scale natural language analyser. Even more significant is the remarkable resemblance in the overall strategy for eliciting information relevant to subsequent language processing programs from the more or less unrestricted natural language text in the definition fields. Still, there are differences in the approaches introduced by specific properties of the machine-readable sources available to the individual projects.

Chodorow et al. follow closely Amsler's prescription for building what is, in essence, a thesaurus: the procedure requires the semantic head of the definition to be located in the description text. The semantic head does not always coincide with the syntactic one, and the group has developed various techniques and heuristics, bringing in constraints from linguistic theory, for finding the genus terms for verbs and nouns. Since the on-line MRD used is W7, and the definitions have been written in a moderately unconstrained style, further techniques had to be developed for disambiguation of the genus terms.

In contrast, Alshawi et al. have taken advantage of a particular property of LDOCE, namely that the definitions in the dictionary are given (when possible) within a limited vocabulary of not more than 2,000 basic words, used *only* in their central meanings. This largely disposes of the need for genus terms disambiguation, and has made it

possible to concentrate on developing robust parsing techniques which are capable of extracting more information from the dictionary and are, in that sense, superior to the approaches proposed by, e.g., Ahlswede (1983), Cowie (1983) and Gaviria (1983). Alshawi's definitions analysis procedure is usually more reliable for precisely locating the semantic heads of definitions, achieves further semantic precision by using other information present in the definition, e.g., modifiers and predication (the result of the analysis process is a frame-like structure with filled-in slots which allows more accurate classification of concepts in a particular domain of discourse), and can be applied not only to verbs and nouns, but to adjectives and adverbs as well.

There are many uses for the semantic information derived in such a way from a machine-readable source. The results of the W7 analysis are going to be incorporated into a semantic component for the CRITIQUE system lexicon (Chodorow et al., 1985). Any program which extracts a taxonomy of concepts from a MRD can be incorporated into an acquisition component of an application system targeted for a specific domain. Thus instead of a module for deriving sortal hierarchies by an interactive (and somewhat stilted) pseudo-natural language dialogue (Haas and Hendrix, 1983), one could imagine a program which would extract the same hierarchy by analysing the definitions of a set of domain-related terms and, as necessary, following chains through the dictionary to establish connecting links.

A program like Alshawi's is particularly useful for applications where a natural language analysis component is being used to interface to a back-end system. In general, approaches to natural language interface design face a range of problems, including limited special vocabulary, large number of synonyms and specialised words which have not been 'pre-programmed' in the interface knowledge base and, in general, difficulties in delimiting discourse domains exactly. The ability to access an on-line MRD, analyse the definition of an unfamiliar word in the input and incorporate the resulting semantic structure in its proper place in the (hierarchical) domain description, thus preventing the system from collapsing with a 'missing vocabulary' message, is invaluable in practical terms.

Alshawi's definitions analyser can be particularly useful in a different context. The definitions analysis program can be applied to the core vocabulary of LDOCE, aiming to derive the taxonomy of defining concepts used by the Longman lexicographers. Such a taxonomy can, quite legitimately, be taken as a guideline for determining a fundamental set of semantic primitives to be used by the interpretation component of a general purpose natural language analyser. Only then will individual word definitions be analysed and corresponding semantic formulae compiled within this primitive vocabulary, using an updated version of Wilks' representation language (Wilks, 1977; Boguraev, 1980), as developed by Carter (1986).

Recent research by Amsler and Walker (personal communication) focuses on an investigation aimed at analysis of compound nouns and more general phrases. Dictionaries and almanacs are used to extract data relevant to these studies. A working hypothesis at this stage of the project, particularly as far as compound nouns are concerned, seems to be that they are instantiations of the hyponyms in their definitions. Within such a framework, on-line access to a semantic hierarchy of defining concepts is a necessary prerequisite for the collection and analysis of empirical data, with the ultimate aim being the derivation of a set of interpretation rules, eventually generalised

through the hierarchy to an extent where they can be used to disambiguate a wide range of non-lexicalised compound phrases. An independent study by Warren (1978) suggests a possible implementation of a program for analysis of complex nominals along very similar lines; indeed, Leonard (1984) describes some very interesting experiments in the computational interpretation of complex noun phrases.

None of the systems and projects mentioned above achieves complete processing of the definitions in a dictionary; nor do they achieve complete semantic coverage. This is due partly to our inability to build fully comprehensive natural language analysers. Of more immediate concern, however, and much more relevant to the subject of this paper, is a different reason for occasional failure, which brings up the question of how reliable is the information to be found in a machine-readable source.

## 5 Reliability and Utility of MRDs

Collective experience makes it feasible, and indeed necessary, to examine some broader aspects of working with machine-readable dictionaries. It seems now possible to ask questions like, for example, what is the overall cost (the manpower effort spent on a project) of attempting to harness what sometimes can be a bulky and unwieldy object? What type of information could reasonably be expected to be found in an MRD? How reliable is this information? What would be the best way to make use of it in any particular context? Eventually, the research community needs to reach a consensus on these topics, together with a feeling for how the range of available dictionaries score on such scale.

There are two factors which are relevant to these issues. The first concerns the information that is likely to be found in a dictionary and raises two further questions. One of these is related to the dictionary coverage, i.e. how representative (non-redundant and yet comprehensive) is it, both within a single entry, and ranging over the entire dictionary. The other concerns its reliability, i.e., do the subfields in the entry conform to prescribed, or expected, format rules; does the content of a field represent adequately and faithfully the facts, as the users understand them to be; and so forth.

The second factor relates directly to how such information is structured and organised, and in particular, how its extraction from the machine-readable source is to be handled. The next section is concerned with issues of the overall packaging of a machine-readable source and I shall postpone questions about organisation and access till then. Below are some more detailed points touching upon the coverage and reliability of data in machine-readable sources likely to be of interest to research in computational linguistics.

The first, and in a way the most obvious, concern is with the expectations for the lexical coverage of the dictionary. Notwithstanding theoretical investigations of the lexicon, and in particular the status of complex lexical items which could either be preinstantiated (e.g., the *full entry model* of the lexicon; see Jackendoff, 1975) or derived on 'as-needed' basis (viz. the *impoverished entry model*), the fact remains that no dictionary can be expected to offer complete coverage of the language. Even if one is designing a customised lexical database, it is necessary to take into account the fact that our vocabulary is finite but open-ended (Meijs, 1986). Thus any attempts to use an existing machine-readable source for a task that requires, or presupposes, dealing

with large bodies of unrestricted samples of text (or speech), leave open the question of what to do when the system strays beyond the boundaries of its lexicon. A number of researchers have pointed out that even though MRDs can be very useful for a range of tasks, it should not be expected that they will be able to supply all the information required by any one of them. Amsler (1984a) emphatically states that "one must admit that there does not exist a word list of English that enumerates the true lexical units of the language" (p.173); Walker and Amsler (1984) present empirical evidence of the huge disparity between the contents of what is, by most standards, a perfectly 'respectable' dictionary (W7), with the vocabulary behind the *New York Times News Service*. Church's work on stress assignment (1985) is partly motivated by a very similar observation.

This insufficient coverage is only partly due to morphological phenomena. As already noted, augmenting the dictionary access software with a powerful morphological analyser, much as it is mandatory, is not going to solve the problem. Still, there is an additional advantage of being able to carry out morphological decomposition accurately. Given the ability to identify redundant words in a source dictionary, it is possible to convert this source into a more compact, as well as as lexically consistent, object. Byrd and Chodorow (1985) reports on a lexical sub-component capable of analysing inflectional affixes, which was used to identify 10,000 redundant entries from a more than 80,000 entry source dictionary. The desire to build a compact, internally consistent and linguistically plausible lexicon motivates the work of Russell et al. (1986): the software developed in the course of the project will be then available to users to create and load, from a machine-readable source, a lexicon, customisable for a range of natural language processing systems (see Boguraev et al., 1986).

Even when the information expected is found in its place in the dictionary, there is still the question of how far we can rely on it. There are many ways in which the client programs of an MRD can be fooled. The most common problem is associated with simple typographic errors. It would seem that the proofreading behind some, not to say most, of the currently available MRDs leaves something to be desired. Mitton (1986) reports on a year long activity aimed at checking, correcting and proofreading the OALDCE tape. Apparently a large number of small errors were introduced in the process of keying the original dictionary data in the computer, and each single one of these would break the accessing programs. Note that this correcting work had to be done before the real use of the dictionary (phonetically guided spelling correction, in this particular case) could commence. Similarly, Yannakoudakis (1983) refers to a series of projects which, after several years running, achieved as a side effect the verification of the *Shorter Oxford Dictionary*. Several research groups in the US (Ed Fox, personal communication) seem to have gone the same route.

Clearly this type of problem could be avoided at source, if a rigorous proofreading procedure is applied by the publisher. Even better, if a tape was used for typesetting as well, the possibility of typographically induced errors would be removed (see the next section). Still, this would not guarantee safe processing: for example, the published version of LDOCE contains about a dozen entries in which parenthesised expressions have a bracket (opening or closing) missing. In an environment where a batch job was submitted overnight to restructure the information on the tape, prior to subsequent use by LISP programs, a considerable amount of time was lost trying to read the tape in memory without losing track of boundaries between individual entries (the generic LISP

'read' function was used, which relies crucially on balanced parentheses).

A different source of unreliability comes from failure of the lexicographers or sub-editors to conform to a consistent format, while working on an entry or a subfield. For example, the 'grammar' employed for the grammatical coding of word senses in LDOCE has not been defined formally anywhere. By and large, a specification of the allowable format of the grammar code fields can be derived by inspecting a sample of LDOCE entries. Leaving aside the fact that this is a time consuming activity, it still carries penalties associated with the fact that the only way to develop such a grammar is by applying it to a large number of entries and manually inspecting the results — a trial-and-error process which could have been altogether avoided by an explicitly defined formal system made available to the lexicographers prior to the dictionary development effort. The same point is made, in a larger context and much more emphatically, by Kazman (1986) who has developed a grammar and associated software for 'parsing' the full entries of the *Oxford English Dictionary*.

Even when seemingly stringent constraints are imposed on dictionary construction, the final product still may suffer from internal inconsistencies. A representative example in this class of problem is the issue of circularity in dictionary definitions. Most publishers adhere to a general principle of reducing complex concepts to a composite of more primitive ones, avoiding straight cases like (the apocryphal) "**recursion**: see *recursion*". The attempt is, where possible, to write the definitions in terms simpler than the words they define. One particular dictionary, namely LDOCE, adopts this maxim wholeheartedly; the preface to the published version claims that "a rigorous set of principles was established to ensure that only the most 'central' meanings of [a controlled vocabulary of] 2,000 words, and only easily understood derivatives, were used" (Procter, 1978, p.ix).

There are several problems here. For a start, no indication is given as to which of the many meanings of the words in this core vocabulary are considered central. Much more serious, and cause for many failings of the definitions analysis programs described above, is the somewhat liberal interpretation of the phrase "only easily understood derivatives". Thus, allowing derivational morphology to creep into the basic set gives some lexicographer the power to use "container" for the definition of "**box<sup>2</sup>**", even though only the verb "contain" is considered to be primitive. Elsewhere, "**container<sup>2</sup>**" is defined as "a very large, usu. metal box...", thus clearly violating the promise of non-circular definitions, described in simpler terms. A program attempting to use the semantic 'hierarchy' derived from analysis of these terms is destined for the equivalent of an infinite loop.

A different kind of trap is introduced by equally liberal use in the definitions of phrasal verbs made up from verbs and particles taken from the restricted vocabulary. For example, the second meaning of "**contain<sup>2</sup>**" is given as "to hold back, keep under control...". While both "hold" (as a verb) and "back" (as a noun) are within the core vocabulary, "hold back", with its own meaning as a phrasal verb, is not. A fundamental assumption on which the definitions analysis program (Alshawi et al., 1985) is based has been violated. Not surprisingly, the result from the conversion of this particular definition into a semantic structure makes no sense and is of no subsequent use.

LDOCE is by no means the only example of an unreliable dictionary source; it is simply the MRD I am most familiar with. A large number of people (Ronald Kaplan,

Mitchell Marcus, Norman Sondheimer, to mention but a few; personal communication, also see Ritchie, 1987) have commented on the significant amount of extra work required in order either to bring other MRDs 'in phase' with their specifications, or to 'patch' a program which attempts to use them computationally.

The current situation then is as follows. On the one hand there is evidence that automatic procedures which are being developed to make use of the information in a dictionary start from the premise that they can rely on what the dictionary specification promises. On the other hand, dictionaries (both machine-readable sources and published versions) fail to fulfil this promise on all accounts. Some obvious questions which follow are: What is the real utility of the programs attempting to derive, digest, compile and further utilise some of the potentially useful data in a machine-readable source? How to establish, for practical purposes, the 'reliability threshold' of such a source, beyond which there is no point in attempting to employ automatic procedures for support of computational linguistics programs?

It is clear that very careful empirical analysis of a dictionary source must be carried out prior to any serious project likely to engage a substantial amount of manpower over a period of time. One way to approach the whole question of making computational sense of a MRD coolheadedly and avoid being badly burnt at the end of the day would involve careful analysis of the available source(s) prior to any investment of human and machine resources. Examples of such analysis are the studies like Akkerman et al.'s (1985) comparative evaluation of the grammatical coding systems for OALDCE and LDOCE, or Moulin et al.'s (1985) analysis of the validity and systematicity of the grammatical information provided by the LDOCE grammar codes, as well as of the consistency of the whole coding system there. An alternative approach to the same problem would be to use an existing source as a starting point for a specific project to do the job better. For instance, a project underway at the University of Amsterdam (Meijs, 1985) essentially attempts, within a particular theoretical linguistics framework, to rectify the situations where the LDOCE word definitions diverge from the standard dictionary practice of maximum economy.

## 6 Structure and Organisation of MRDs

An important question, already alluded to in this paper on several occasions, concerns the organisation and structuring of the vast amounts of information available in dictionaries. The issue is not simply one of maintaining large data files. On the one hand there is an inherent contradiction in the nature of a dictionary: it carries too much free text to be easily amenable to conventional database management techniques and yet has a lot more structure than could be comfortably accommodated by an information retrieval system. On the other hand, given the growing tendencies, respectively, in the publishers' camp to apply computer-based techniques to the task of dictionary production (see, e.g., the recent work on computerising the OED, or the similar, though on a more modest scale, project undertaken to make the *Chambers 20th Century Dictionary of English* available on-line), and in the computational linguists' camp to make serious use of the resulting machine-readable dictionary sources, it is only natural to ask where the optimal meeting point of two alternative ways of delivering a dictionary should be.

So far computational linguists have had to do with computer tapes, typically typesetting ones, where the organisation of data has followed certain lexicographic and typographic principles and the primary underlying assumption in the data gathering and structuring process has been that a dictionary is a printed object to be used by an intelligent human. Consequently, a tape could be viewed, in the abstract, as a character stream, providing a mixture of typesetting commands and real data. Any program of the kind discussed in this paper, which wishes to make use of the information content of a MRD, must tackle the problem of extracting this information from the character stream and (re)structuring it appropriately to its purposes. Note that this is not a trivial task, as lexicographic data are often conveyed by form, in addition to content, in the printed version of a dictionary — thus the typesetting codes on the computer tape cannot simply be ignored.

If the same dictionary is going to be made available to a computer program, necessarily with different characteristics and objectives, perhaps its underlying, core format needs revising? There seem to be strong arguments, from the point of view of computational linguistics practices, in favour of applying database design and management techniques to a computerised dictionary, and loading it into an appropriately structured database. This database could subsequently be used both for generating a typesetting tape and for supporting a range of applications similar to the ones discussed in this paper. The management of dictionary data in such 'meta'-format would speed up considerably the process of tailoring the machine-readable source to a particular task. In addition, it is a much better option for the computational linguist who now need not take a computerised dictionary in the form of a character stream requiring a non-trivial amount of comparatively low-level, but error-prone and very time-consuming, preprocessing and restructuring of the data before they are in a form suitable for use by any specific program.

Some dictionary publishers have long been aware of the advantages of having a 'master' source on tape, particularly if it has some structure imposed on it. Thus Longman have used their three major sources, the *Longman Dictionary of the English Language* (LDEL), LDOCE and *Roget's Thesaurus*, all available on-line, to spawn new titles electronically. Chambers are in the process of structuring their *Chambers' Dictionary of 20th Century English* (Kirkpatrick, 1983) with the view of putting it on-line. Oxford University Press has launched a mammoth project not only to key in the complete text of the *Oxford English Dictionary*, but to design a special-purpose database best suited for the vast range of tasks elicited by a survey amongst the potential users of the electronic version of the OED (see Lesk, 1986a, 1986b; Stubbs 1986; Tompa, 1986).

Nonetheless, the problem is not simply reduced to the design of, say, a relational database. Just as an encyclopedia is useless unless the information in it is divided and assembled into natural units, a database is useless unless it is organised in a way that renders accessible the information a user may require. Michiels (1983) makes a distinction between a machine-readable and a computerised dictionary; the former being simply a dictionary available in machine-readable form, and the latter carrying some structure with it. He goes on to point out that the typesetting tape for LDOCE is the only truly computerised dictionary of English. Even so, converting the tape to a database suitable for a wide range of users — lexicographers and computational linguists alike — would be far from trivial, precisely because the structure imposed on the tape has

not evolved on the basis of considerations of the classes of task; particularly within the framework of computational linguistics research that the dictionary might be applied to.

Johnson (1985) makes a similar point: simply trying to apply a large database management system (DBMS) to an application much different from what the original designers had in mind would fail to meet criteria of flexibility and extensibility. In particular, clients of a dictionary database are likely to regard such an artifact as a dynamic object. They are not only going to ask questions whose answers are to be found directly and explicitly in the database (e.g., "select words derived from old German, which have entered English between the years of 1650 and 1700"). Computational linguists are likely to pose queries requiring non-trivial, on-line, processing of the lexical information in the database: examples here would include extraction of words which are 'close' to a given word along dimensions such as spelling, meaning, or sound (see Byrd, 1983), or enumeration of verbs with specified syntactic subcategorisation patterns (Alshawi et al., 1985).

This paper has already discussed classes of possible uses for the information to be found in a machine-readable source. In the general case, whenever a dictionary is to support research in computational linguistics (e.g., as discussed earlier, in the area of speech analysis, for grammar development, or for elicitation of a basic vocabulary set of primitives), as opposed simply to provide selected data for a particular task (e.g., construction of a word-list indexed to a grammar or a lexicon to be used by a spelling checker program), a straightforward mapping into a database is not going to be enough. Calzolari (1984b) points out that the object which is ultimately going to be invaluable for computational linguistics research is not the machine-readable dictionary itself, but an extension of a 'relational' database, with complex interlinked structures superimposed on the individual entries, and with the particular properties of being multi-functional and multi-perspective. Domenig (1986) similarly rejects the general purpose database management system as both too powerful and not powerful enough for the task of multi-perspective access to potentially dynamic information in a lexical database. The 'Lexical Project' at the Center for Study of Language and Information (CSLI Monthly, 1986) is also investigating the conceptual and computational issues related to compiling the lexical knowledge required by a host of syntactic and semantic frameworks currently under development at CSLI into a single on-line lexicon.

Byrd (1986) offers further arguments in support of a flexible organisation of an on-line dictionary. He points out the advantages of, and indeed the need for, a number of interfaces allowing easy navigation through the different kinds of data, which would be required by two qualitatively different classes of clients in the automated office: human users and computer programs.

The details of the issue of what would be a suitable meta-format for a dictionary entry, allowing various classes of users easy and transparent access to selected aspects of the data, together with the related question of arriving at some consensus on the functional properties of a database capable of supporting a wide range of applications, are still open research.

## 7 Conclusion

This paper has looked at a number of contexts within computational linguistics, to which the use of a machine-readable dictionary is of great relevance. It is clear that many language processing applications are seeking lexical and knowledge support from suitable, or available, on-line sources. The range of open questions, arising from the widely differing requirements of individual systems, as well as the non-uniform modes of use of MRDs, demonstrates that this is an active field of research, holding promise for practical natural language processing. One thing, however, is clear now. In an environment where very heavy use is made of artifacts primarily designed to be read by humans, and not used by computers, we are pushing the limits of effective and reliable utilisation of machine-readable sources. The next generation of dictionaries must be designed from new principles, and delivered in a different form; the only way to do this properly is to take into account the requirements of both people and programs, as well as the current state of knowledge in lexicography, computational linguistics, language engineering, information structuring and database management systems.

## Acknowledgments

I would like to thank Hiyan Alshawi, David Carter, Karen Sparck Jones and in particular Archibal Michiels for their comments on draft versions of this paper. Some of the points raised here can be traced back to the five days of discussions at a workshop on *Automating the Lexicon*. The work was carried out during the author's tenure of an Advanced Research Fellowship at the University of Cambridge Computer Laboratory, and administered by the UK Science and Engineering Research Council.

## References

- [1] Ahlswede, Thomas., "A Linguistic String Grammar of Adjective Definitions from Webster's Seventh Collegiate Dictionary", MS Thesis, Illinois Institute of Technology, Chicago, Illinois, 1983.
- [2] Akkerman, Erik; Masereuw, Pieter; and Meijls, Willem, "Designing a Computerised Lexicon for Linguistic Purposes", ASCOT Report No. 1, CIP-Gegevens Koninklijke Bibliotheek, Den Haag, The Netherlands, 1985.
- [3] Alshawi, Hiyan, "Processing Dictionary Definitions with Phrasal Pattern Hierarchies", in B. Boguraev and E. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*, Longman, Harlow and London, 1989.
- [4] Alshawi, Hiyan; Boguraev, Branimir; and Briscoe, Ted, "Towards a Lexicon Support Environment for Real Time Parsing", *Proceedings of the Second Conference of the European Chapter of the Association for Computational Linguistics*, pp.171–178, Geneva, 1985.
- [5] Amsler, Robert, "The Structure of the Merriam-Webster Pocket Dictionary", Ph.D. Thesis, University of Texas, Austin, Texas, 1980.

- [6] Amsler, Robert, "Experimental Research on Knowledge Representation for Lexical Disambiguation of Full-Text Sources", Research Proposal for the NSF, SRI International, Menlo Park, California, 1983.
- [7] Amsler, Robert, "Machine-Readable Dictionaries", in Martha E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST)*, published by the American Society for Information Science, 19, 1984a.
- [8] Amsler, Robert, "Lexical Knowledge Bases", panel session on Machine-Readable Dictionaries, in *Proceedings of the Tenth International Congress on Computational Linguistics*, pp.458–459, Stanford, CA, 1984b.
- [9] Atwell, Eric; Leech, Geoffrey; and Garside, Roger, J. Aarts and W. Meijs (Eds.), "Analysis of the LOB Corpus: Progress and Prospects", in *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*, Rodopi (Costerus. New Series; Vol. 45), Amsterdam, 1984.
- [10] Ballard, Bruce and Biermann, Alan, "Workshop on Transportable Natural Language Interfaces", Duke University, October, 1984, presentations published in a special issue of *ACM Transactions on Office Information Systems*, April, 1985.
- [11] Bell, Colin and Jones, Kevin, "Back-of-the-Book Indexing: A Case for the Application of Artificial Intelligence", in M. MacCaffery and K. Gray (Eds.), *Informatics 5: The Analysis of Meaning*, Proceedings of a conference held by the Aslib Informatics Group and the BCS Information Information Retrieval Specialist Group, Oxford, UK, March 1979.
- [12] Bobrow, Robert and Webber, Bonnie, "Knowledge Representation for Syntactic/Semantic Processing", in *Proceedings of the First National Conference on Artificial Intelligence*, pp. 316–323, Stanford, California, 1980.
- [13] Boguraev, Branimir, "Automatic Resolution of Linguistic Ambiguities", Ph.D. Dissertation, also available as Technical Report No.11, Computer Laboratory, University of Cambridge, Cambridge, 1980.
- [14] Boguraev, Branimir; Carroll, John; Pulman, Steve; Russell, Graham; Ritchie, Graeme; Black, Alan; Briscoe, Ted; and Grover, Claire, "The Lexical Component of a Natural Language Toolkit", paper presented at *An International Workshop on Automating the Lexicon*, Grosseto, Italy, 1986.
- [15] Boguraev, Branimir; Copestake, Ann; and Sparck Jones, Karen, "Inference in Natural Language Front Ends for Databases", in E. Meersman, and J. Sowa (Eds.), *Knowledge and Data: Proceedings of IFIP WG 2.6 Working Conference* North-Holland, Amsterdam, 1988.
- [16] Boguraev, Branimir and Sparck Jones, Karen, "How to Drive a Database Front End Using General Semantic Information", in *Proceedings of a Conference on Applied Natural Language Processing*, Santa Monica, California, pp.:81–89, 1983.

- [17] Boguraev, Branimir and Briscoe, Ted, "Large Lexicons for Natural Language Processing: Exploiting the Grammar Coding System of LDOCE". *Computational Linguistics*, 13 (3–4), 1987.
- [18] Brachman, Ronald and Schmolze, James, "An Overview of the KL-ONE Knowledge Representation System", *Cognitive Science*, 9 (2) pp. 171–216, 1985.
- [19] Briscoe, Ted, "Report of a Dictionary Syndicate", in *Proceedings of an Alvey Speech Club Workshop*, Warwick University, July 1985.
- [20] Byrd, Roy, "Word Formation in Natural Language Processing Systems", in *Proceedings of Eighth International Joint Conference on Artificial Intelligence*, pp. 704–706, Karlsruhe, Germany, 1983.
- [21] Byrd, Roy, "Dictionary Systems for Office Practice", paper presented at *An International Workshop on Automating the Lexicon*, Grosseto, Italy, 1986.
- [22] Byrd, Roy and Chodorow, Martin, "Using an On-line Dictionary to Find Rhyming Words and Pronunciations for Unknown Words", in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pp.277–283, Chicago, Illinois, 1985.
- [23] CSLI Monthly, "Report of Research Activities", a monthly publication of the Center for the Study of Language and Information, 1 (1), Stanford University, Stanford, California, 1986.
- [24] Calzolari, Nicoletta, "Lexical Definitions in a Computerised Dictionary", *Computers and Artificial Intelligence*, 2 (3), 1983.
- [25] Calzolari, Nicoletta, "Detecting Patterns in a Lexical Database", in *Proceedings of the Tenth International Congress on Computational Linguistics*, pp.170–173, Stanford, California, 1984a.
- [26] Calzolari, Nicoletta, "Machine-Readable Dictionaries, Lexical Data Bases and the Lexical System", panel session on Machine-Readable Dictionaries, *Proceedings of the Tenth International Conference on Computational Linguistics*, p.460, Stanford, California, 1984b.
- [27] Calzolari, Nicoletta and Picchi, Eugenio, "The Machine-Readable Dictionary as a Powerful Tool for Consulting Large Textual Archives", in *Papers from Conference on Automatic Processing of Art History Data and Documents*, 1, 277–288, Pisa, 1984.
- [28] Carter, David, "A Shallow Processing Approach to Anaphor Resolution", Technical Report No. 88, Computer Laboratory, Cambridge University, Cambridge, 1986 (also published by Ellis Horwood, Chichester, England, 1988).
- [29] Cater, Arthur, "Conceptual Primitives and Their Metaphorical Relationships", in E. Reilly (Ed.), *Communication Failure in Dialogue*, North Holland, Amsterdam, 1987.

- [30] Chodorow, Martin; Byrd, Roy; and Heidorn, George, “Extracting Semantic Hierarchies from a Large On-Line Dictionary”, in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pp.299–304, Chicago, Illinois, 1985.
- [31] Church, Kenneth, “Stress Assignment in Letter to Sound Rules for Speech Synthesis”, in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pp.246–253, Chicago, Illinois, 1985.
- [32] Cottrell, Garrison and Small, Steven, “A Connectionist Scheme for Modelling Word Sense Disambiguation”, *Cognition and Brain Theory*, 6 (1), 89–120, 1983.
- [33] Cowie, James, “Automatic Analysis of Descriptive Texts”, in *Proceedings of Conference on Applied Natural Language Processing*, pp.117–123, Santa Monica, California, 1983.
- [34] Cumming, Susanna, “A Guide to Lexical Acquisition in the JANUS System”, ISI Research Report ISI/RR-85-162, Information Sciences Institute, Marina del Rey, California, 1986a.
- [35] Cumming, Susanna, “The Distribution of Lexical Information in Text Generation”, in this volume.
- [36] Domenig, Marc and Shann, Patrick, “Towards a Dedicated Database Management System for Dictionaries”, in *Proceedings of the 11th International Conference on Computational Linguistics*, pp.91–96, Bonn, Germany, 1986.
- [37] Fawthrop, David and Yannakoudakis, E.J., “An Intelligent Spelling Error Corrector”, *Information Processing and Management*, 19 (2), 101–108, 1983.
- [38] Garside, Roger and Geoffrey Leech, “Grammatical Tagging of the LOB Corpus: General Survey”, in S. Johansson (Ed.), *Computer Corpora in English Language Research*, Norwegian Computing Centre for the Humanities, Bergen, 1982.
- [39] Gaviria, Gustavo, “Une approche pour amorcer le processus de comprehension et d'utilisation du sens des mots en langage naturel”, Centre Nacional de la Recherche Scientifique, Paris, 1983.
- [40] Gazdar, Gerald, Klein, Ewan, Pullum, Geoffrey and Sag, Ivan, *Generalized Phrase Structure Grammar*, Basil Blackwell Publisher Ltd., Oxford, 1985.
- [41] Grosz, Barbara and Stickel, Mark, “Research on Interactive Acquisition and Use of Knowledge”, Final Report, SRI International, Menlo Park, California, 1983.
- [42] Haas, Norman and Hendrix, Gary, “Learning by Being Told: Acquiring Knowledge for Information Management”, in R. Michalski, J. Carbonell and T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga Publishing Company, Palo Alto, California, 1983.
- [43] Hanks, Patrick, *The Collins English Dictionary*, William Collins Sons and Co. Ltd., Glasgow, 1979.

- [44] Heidorn, George; Jensen, Karen; Miller, Lance; Byrd, Roy; and Chodorow, Martin, “The EPISTLE Text-Critiquing System”, *IBM Systems Journal*, 21 (3), 305–326, 1982.
- [45] Hirst, Graeme, *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, Cambridge, 1987.
- [46] Hornby, A.S., *Oxford Advanced Learner’s Dictionary of Current English*, Third Edition, Oxford University Press, Oxford, 1974.
- [47] Huttenlocher, Daniel and Zue, V, “Phonotactic and Lexical Constraints in Speech Recognition”, *Proceedings of the National Conference on Artificial Intelligence*, pp.172–176, Washington, D.C., 1983.
- [48] Ingria, Robert, “Lexical Information for Parsing Systems: Points of Convergence and Divergence”, in this volume.
- [49] Jackendoff, Ray, “Semantic and Morphological Regularities in the Lexicon”, *Language*, 51, 639–671, 1975.
- [50] Jackendoff, Ray and Grimshaw, Jane, “A Key to the Verb Catalog”, unpublished mimeo, under NSF Grant IST-84-20073, Information Structure of a Natural Language Lexicon, Program in Linguistics and Cognitive Science, Brandeis University, Waltham, MA, 1985.
- [51] Johnson, Mark, “Computer Aids for Comparative Dictionaries”, *Linguistics*, 232, 285–302, 1985.
- [52] Johnson, Tim, *Natural Language Computing: The Commercial Applications*, Ovum Press Ltd., London, 1985, (revised and updated edition as *Natural Language Markets: Commercial Strategies*, Ovum Press, 1990).
- [53] Kaplan, Ronald and Bresnan, Joan, “Lexical-Functional Grammar: A Formal System for Grammatical Representation”, in J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, Massachusetts, 1982.
- [54] Katz, Jerrold and Fodor, Jerry, “The Structure of a Semantic Theory”, *Language*, 39, 170–210, 1963.
- [55] Kay, Martin, “The Dictionary Server”, panel session on Machine-Readable Dictionaries, in *Proceedings of the Tenth International Conference on Computational Linguistics*, p.461, Stanford, California, 1984a.
- [56] Kay, Martin, “Functional Unification Grammar: A Formalism for Machine Translation”, in *Proceedings of the Tenth International Congress on Computational Linguistics*, pp.75–79, Stanford, California, 1984b.
- [57] Kay, Martin and Kaplan, Ronald, “Phonological Rules and Finite State Transducers”, paper presented at the Annual Meeting of the Association for Computational Linguistics, New York City, 1981.

- [58] Kazman, Rick, “Structuring the Text of the Oxford English Dictionary through Finite State Transduction”, Master’s Thesis (Department of Computer Science Technical Report TR-86-20), University of Waterloo, Waterloo, Ontario, 1986.
- [59] Kirkpatrick, E. (Ed.), *Chambers 20th Century Dictionary of English*, New Edition, W&R Chambers Ltd., Edinburgh, 1983.
- [60] Klingbiel, Paul, “Phrase Structure Rewrite Systems in Information Retrieval”, *Information Processing and Management*, 21 (2), 113–126, 1985.
- [61] Koskenniemi, Kimmo, “Two-level Morphology: a General Computational Model for Word-Form Recognition and Production”, Publication No. 11, University of Helsinki, Finland, 1983.
- [62] LDEL, *Longman Dictionary of the English Language*, Longman Group Limited, Harlow, 1984.
- [63] Lenat, Douglas; Prakash, Mayank; and Shepherd, Mary, “CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks”, MCC Nonproprietary Technical Report AI-055-85, 1985 (also in *AI Magazine*, 6 (4), 65–92, Winter 1986).
- [64] Leonard, Rosemary, *The Interpretation of English Noun Sequences on the Computer*, North-Holland, Amsterdam, 1984.
- [65] Lesk, Michael, “Information in Data: Using the Oxford English Dictionary on a Computer”, summary of a Conference on Information in Data held in the Centre for the New OED, University of Waterloo, November 1985; also in *ACM SIGIR Forum*, 20 (12), Spring 1986a.
- [66] Lesk, Michael, “Why I Want the OED on My Computer, and When I’m Likely to Have It”, *SIGCUE Newsletter*, 2nd Quarter, 1986b.
- [67] McArthur, Tom, *Longman Lexicon of Contemporary English*, Longman Group Limited, Harlow, 1981.
- [68] McDermott, Drew, “Artificial Intelligence Meets Natural Stupidity”, *SIGART Newsletter*, 57, April 1976; also reprinted in John Haugeland (Ed.), *Mind Design*, MIT Press, Cambridge, Massachusetts, 1981.
- [69] Mark, William, “Representation and Inference in the CONSUL System”, in *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp.375–381, Vancouver, British Columbia, 1981.
- [70] Martin, Paul; Appelt, Douglas; and Fernando Pereira, “Transportability and Generality in a Natural Language Interface System”, in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp.573–581, Karlsruhe, Germany, 1983.

- [71] Masterman, Margaret; Needham, Roger; Sparck Jones, Karen; and Mayoh, Bill, “AGRICOLA INCURVO TERRAM DIMOVIT ARATRO’. First Stage Translation into English with the Aid of Roget’s Thesaurus”, Report (ML84) ML92, Cambridge Language Research Unit, Cambridge, 1957.
- [72] Meijls, Willem, “Linguistically Useable Meaning Characterisations (LINKS) in the Lexicon”, Project Description, English Department, University of Amsterdam, Amsterdam, 1985.
- [73] Meijls, Willem, “Lexical Organisation from Three Different Angles”, *Journal of the Association of Literary and Linguistic Computing*, 13 (1), 1986.
- [74] Michiels, Archibal and Noel, Jacques, “Approaches to Thesaurus Production”, *Proceedings of the 9th International Congress on Computational Linguistics*, Prague, Czechoslovakia, 1982.
- [75] Michiels, Archibal, “Exploiting a Large Dictionary Data Base”, Doctoral Dissertation, University of Liege, Liege, 1982.
- [76] Michiels, Archibal, “Automatic Analysis of Texts”, in *Informatics 7*, Proceedings of a conference held by the Aslib Informatics Group and the Information Retrieval Group of the British Computer Society, pp.103–120, Cambridge, 1983.
- [77] Miller, George, “WordNet: A Dictionary Browser”, paper presented at the Conference on Information in Data, University of Waterloo Centre for the New OED, Ontario, Canada, 1985 (see also a special issue of *International Journal of Lexicography*, 3 (4), 1990).
- [78] Mitton, Roger, “A Partial Dictionary of English in Computer Usable Form”, Mimeo, Computer Science Department, Birkbeck College, University of London, 1986.
- [79] Moulin, André; Jaques Jansen and Archibal Michiels, “Computer Exploitation of LDOCE’s Grammatical Codes”, paper presented at a Conference on Survey of English Language, Lund, 1985.
- [80] MPD, *New Merriam-Webster Pocket Dictionary*, Pocket Books, New York, 1964.
- [81] Pollock, Joseph, “Spelling Error Detection and Correction by Computer: Some Notes and Bibliography”, *Journal of Documentation*, 38 (4), 282–291, 1982.
- [82] Pollock, Joseph, and Zamora, Antonio, “Automatic Spelling Correction in Scientific and Scholarly Text”, *Communications of the ACM*, 27 (4), 358–368, 1984.
- [83] Procter, Paul, *Longman Dictionary of Contemporary English*, Longman Group Limited, Harlow and London, 1978.
- [84] Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey and Svartvik, Jan, *A Contemporary Grammar of English*, Longman Group Limited, London, 1972.

- [85] Rich, Elaine, “Natural-Language Interfaces”, *IEEE Computer*, 39–47, September, 1984.
- [86] Ritchie, Graeme, “Discussion Session on the Lexicon”, in Whitelock et al., 1987.
- [87] Ritchie, Graeme; Black, Alan; Pulman, Steve; and Russell, Graham, “The Edinburgh/Cambridge Morphological Analyser and Dictionary System: User Manual”, Software Paper 10, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, 1987.
- [88] Russell, Graham; Pulman, Steve; Ritchie, Graeme; and Black, Alan, “A Computational Framework for Lexical Description”, *Computational Linguistics*, 13 (3–4), 290–307, 1987.
- [89] Shieber, Stuart, “Criteria for Designing Computer Facilities for Linguistic Analysis”, *Linguistics*, 232, 189–211, 1985.
- [90] Shipman, David and Zue, Victor, “Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems”, *Conference Record*, pp.546–549, IEEE International Conference on Speech Acoustics and Signal Processing, Paris, 1982.
- [91] Small, Steven, “Word Expert Parsing: a Theory of Distributed Word-Based Natural Language Understanding”, Ph.D. Dissertation, also available as Technical Report TR-954 / NSG-7253, Department of Computer Science, University of Maryland, 1980.
- [92] Sowa, John and May, Eileen, “Implementing a Semantic Interpreter Using Conceptual Graphs”, *IBM Journal of Research and Development*, 30 (1), 57–69, 1986.
- [93] Streeter, Lynn A., “The Acoustic Determination of Phrase Boundary Perception”, *Journal of Acoustic Society of America*, 64 (6), 15–82, 1978.
- [94] Stubbs, John, “A Database-in-Waiting: The OED Becomes the New OED”, paper presented to the Computers and the Humanities Conference, University of Toronto, Toronto, 1986.
- [95] Thompson, Henry, “Natural Language Processing: A Critical Analysis of the Structure of the Field, with Some Implications for Parsing”, in K. Sparck Jones and Y. Wilks (Eds.), *Automatic Natural Language Parsing*, Ellis Horwood, Chichester, 1983.
- [96] Thorndike, E.L. and Lorge, I., *The Teachers Word Book of 30,000 Words*, Teachers College Press, Teachers College, Columbia University, New York, 1944.
- [97] Tompa, Frank, “Database Design for a Dictionary of the Future”, unpublished manuscript, Centre for the New Oxford English Dictionary, University of Waterloo, Waterloo, Ontario, 1986.

- [98] Tucker, Allen and Nirenburg, Sergei, "Machine Translation", in Martha E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST)*, American Society for Information Science, 19, 1984.
- [99] Walker, Donald and Amsler, Robert, "The Use of Machine-Readable Dictionaries in Sublanguage Analysis", in Richard Kittredge (Ed.), *Proceedings of Workshop on Sublanguage Analysis*; New York, 1984 (also available in *Analyzing Language in Restricted Domains*, (Grishman, R. and Kittredge, R. (Eds.), Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986).
- [100] Waltz, David, "Artificial Intelligence: An Assessment of the State-of-the-Art and Recommendation for Future Directions", *AI Magazine*, Fall 1983.
- [101] Waltz, David and Pollack, Jordan, "Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation", *Cognitive Science*, 9 (1), 1985.
- [102] Warren, Beatrice, *Semantic Patterns of Noun-Noun Compounds*, Gothenburg Studies of English 41, Goteborg: Acta Universitatis Gothenburgensis, 1978.
- [103] Webster's Seventh New Collegiate Dictionary, C.&C. Merriam Company, Springfield, Massachusetts, 1967.
- [104] Wehrli, Eric, "Design and Implementation of a Lexical Data Base", *Proceedings of the Second Conference of the European Chapter of the Association for Computational Linguistics*, pp.146–153, Geneva, 1985.
- [105] Whitelock, Peter; Somers, Harry; Bennett, Paul; Johnson, Rod; and Wood, Mary (Eds.), *Linguistic Theory and Computer Applications*, Academic Press, New York, 1987.
- [106] Wilks, Yorick, "Good and Bad Arguments for Semantic Primitives", *Communication and Cognition*, 10, 181–221, 1977.
- [107] Yannakoudakis, E.J., "Expert Spelling Errors Analysis and Correction", in *Informatics 7*, Proceedings of a conference held by the Aslib Informatics Group and the Information Retrieval Group of the British Computer Society, Cambridge, 1983.

## Appendix

Since the Grosseto workshop on Automating the Lexicon, two separate — and yet closely related — developments can be observed. On the one hand, both the theoretical linguistics and the computational linguistics communities have become much more aware of the tremendous potential that machine-readable dictionaries hold: for large-scale lexicon extraction, for better understanding of the nature of the lexicon, and for verification of linguistic hypotheses and theories. Representative examples here may be found in the work of Boguraev & Briscoe (1987), Neff & McCord (1990), Pustejovsky (1991), Levin (1991), and Levin (1993); also see Evens (1989) for a more recent review of activities in the area of machine-readable dictionaries and their use in lexical research, as well as Briscoe (1991) for a detailed discussion of the interaction between lexical issues and acquisition methodologies. On the other hand, return to empiricism in the field of natural language processing has led to an increased activity in the area of corpus-based language studies, and in particular, corpus-driven acquisition of lexical information: see, for instance, Church (1985), Atkins (1987), Church et al. (1991), Hindle (1990), Justeson & Katz (1991); most recently, this line of work has been best illustrated by selected papers in Armstrong (1993), and Boguraev & Pustejovsky (1993); for in-depth review, see Basili et al. (1994).

In addition to pursuing its central goal —semi-automatic derivation of facts about words, word meanings, and word uses— computational lexicology has elaborated a range of other equally important questions. For instance, some of the primary concerns of the field to date include: developing methods and techniques for structuring and analysis of on-line lexical resources (see, for instance, Amsler & Tompa, 1988; Neff & Boguraev, 1989; Boguraev et al., 1990); looking for clues to the structure and organisation of the human lexicon (Miller, 1990; Aitchison, 1987); and defining methodologies for lexical semantics research (Calzolari & Picchi, 1988; Copestake, 1990; Zernik, 1991; Boguraev, 1991b, offer a representative sample of work under this heading). An especially pertinent set of questions, at the core of computational analysis of language on the basis of information available in dictionaries (as well as in other types of machine-readable resources), concerns the relationship between natural language processing, formal syntax and lexical semantic theories, and the way in which this relationship is reflected in the kind of information sought in dictionaries for incorporation into a computational lexicon. The issues here are at least two-fold: what would constitute appropriate mining and filtering procedures for identifying computationally relevant lexical information in sources which are, at best, not explicit with respect to that information (see, for instance, Boguraev & Briscoe, 1989, and Boguraev & Pustejovsky, 1993); and what is the nature of the interaction between lexical semantics, knowledge representation, and natural language processing (Pustejovsky & Bergler, 1992, Pustejovsky & Boguraev, 1993).

To a large extent, recent work reflects a change in view concerning the predominant paradigm of computational lexicology: whereas early efforts for utilising dictionary data were aimed primarily at what had been explicitly stated in the dictionary entries (as the majority of position papers at the Grosseto meeting exemplify), more recent developments have focussed on carrying out much more detailed, and global, analysis of the sources with a view of uncovering information which turns out to be systematically encoded *across* the entire source(s). Thus, examples of extraction processes in

earlier frameworks include: acquisition of information about control and logical type of predicates (Boguraev & Briscoe, 1987), extraction of semantic features (e.g. selectional restrictions) for lexical disambiguation (Byrd et al., 1987), or derivation of information about stress assignment (Church, 1985). In contrast, the desire to make maximal use of the information in dictionaries has promoted work exploiting the distributed nature of the lexical knowledge encoded in these sources; representative examples here include: developing better models of speech recogniser front ends (Carter, 1989), building semantically sound lexical hierarchies (Copestake, 1990), sprouting networks of lexical relations between words (Chodorow et al., 1988), deriving empirical evidence for the existence of semantically coherent word clusters (Wilks et al., 1993), refining such networks to reflect word-sense distinctions (Guthrie et al., 1990), and even extending such activities beyond the boundaries of a single dictionary (Byrd, 1989). A number of related issues, largely to do with populating richer lexical structures which introduce an additional dimension to the notion of lexical relation and account for the permeability among word senses are discussed in Boguraev (1991b). In general, there is an observable shift of emphasis, from extracting primarily syntactic properties of words, to seeking and formalising lexical semantic information (Briscoe et al., 1993, is a particularly rich collection, covering a range of techniques of lexical representation and their use for maintaining lexicons extracted semi-automatically from machine-readable dictionaries).

At least one characteristic of recent research is the tendency to examine critically the notion of building a computational lexicon on the basis of existing machine-readable dictionaries; particularly common is a certain amount of scepticism toward attempts to fully instantiate such a lexicon by automatic means. Still, there is a shared attitude that while there are many ways in which dictionaries might fail as sources, there are also ways to maximise the value of information found in them; Boguraev (1991a) presents a number of strategies developed for this purpose. Moreover, dictionary sources were just the beginning; in addition to specific lexical properties not covered by the process of human lexicography, there is a large number of different word classes which remain outside of the coverage of any such source. Consequently, techniques and methods for lexicon acquisition from text materials are being added to the inventory of tools and methodologies for computational lexicology (Boguraev & Pustejovsky, 1994).

## References

- [1] Aitchison, J., *Words in the Mind: An Introduction to the Mental Lexicon*, Basil Blackwell, Oxford, UK, 1987.
- [2] Amsler, R. and W. Tompa, “An SGML-Based Standard for English Monolingual Dictionaries”, *Information in Text: Proceedings of the Fourth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Waterloo, Ontario, pp.61–80, 1988.
- [3] Armstrong, S.(ed.), Special Issue on Using Large Corpora, *Computational Linguistics*, 19(1-2), 1993.
- [4] Atkins, B., “Semantic ID Tags: Corpus Evidence for Dictionary Senses”, *Proceedings of the Third Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Waterloo, Ontario, 17–36, 1987.
- [5] Basili, R., M.-T. Pazienza and P. Velardi, *Lexicon Acquisition for Real Natural Language Processing Systems*, Cambridge University Press, Cambridge (forthcoming), 1994.
- [6] Boguraev, B., “Building a Lexicon: The Contribution of Computers”, in B. Boguraev (Ed.), Special issue on computational lexicons, *International Journal of Lexicography*, 4(3), 1991a.
- [7] Boguraev, B. (ed.), Special Issue on Computational Lexicons, *International Journal of Lexicography*, 4(3), 1991b.
- [8] Boguraev, B. and E. Briscoe, “Large Lexicons for Natural Language Processing: Exploiting the Grammar Coding System of LDOCE”, *Computational Linguistics*, 13(3-4) 203–18, 1987.
- [9] Boguraev, B. and E. Briscoe, *Computational Lexicography for Natural Language Processing*, Longman Limited, Harlow and London, 1989.
- [10] Boguraev, B., T. Briscoe, J. Carroll and A. Copestake, “Database Models for Computational Lexicography”, *Proceedings of the Fourth International Congress of the European Association from Lexicography (EURALEX-VOX)*, Malaga, Spain, published by Biblograf, Barcelona, 1992.
- [11] Boguraev, B. and J. Pustejovsky (eds.), “Acquisition of Lexical Knowledge from Texts”, Proceedings of the Second International SIGLEX Workshop, Columbus, Ohio, 1993.
- [12] Boguraev, B. and J. Pustejovsky (eds.), *Corpus Processing for Lexicon Acquisition*, MIT Press, Cambridge, Mass. (forthcoming), 1994.
- [13] Briscoe, E., “Lexical Issues in Natural Language Processing”, in E. Klein and F. Veltman (Eds.), *Natural Language and Speech*, Springer-Verlag, Berlin, pp.39–68, 1991.

- [14] Briscoe, E., V. de Paiva, and A. Copestake, *Inheritance, Defaults, and the Lexicon*, Cambridge University Press, Cambridge, 1993.
- [15] Byrd, R., “Discovering Relationships Among Word Senses”, *Proceedings of the Third Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Oxford, pp.67–80, 1989.
- [16] Byrd, R., N. Calzolari, M. Chodorow, J. Klavans, M. Neff and O. Rizk, “Tools and Methods for Computational Lexicology”, *Computational Linguistics*, 13(3–4), 219–40, 1987.
- [17] Calzolari, N. and E. Picchi, “Acquisition of Semantic Information from an On-Line Dictionary” *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary, pp.87–92, 1988.
- [18] Carter, D., “LDOCE and Speech Recognition”, in B. Boguraev and E. Briscoe (eds.), *Computational Lexicography for Natural Language Processing*, Longman, London and New York, pp.135–52, 1989.
- [19] Chodorow, M., Y. Ravin and H. Sachar, “A Tool for Investigating the Synonymy Relation in a Sense Disambiguated Thesaurus”, *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, TX, pp.144–51, 1988.
- [20] Church, K., “Stress Assignment in Letter-to-Sound Rules for Speech Synthesis”, *Proceedings of the 23rd Annual Meeting of the ACL*, Chicago, pp. 246–54, 1985.
- [21] Church, K., W. Gale, P. Hanks and D. Hindle, “Using Statistics in Lexical Analysis”, in U. Zernik (Ed.), *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.
- [22] Copestake, A., “An Approach to Building the Hierarchical Element of a Lexical Knowledge Base from a Machine Readable Dictionary”, in *Proceedings of the First International Workshop on Inheritance in Natural Language Processing*, Tilburg, The Netherlands, pp.19–29, 1990.
- [23] Evans, M., “Computer-Readable Dictionaries”, in M. Williams (Ed.), *Annual Review of Information Science and Technology*, vol. 24, Elsevier Science Publishing Co., The Netherlands, 1989.
- [24] Guthrie, L., B. Slator, Y. Wilks and R. Bruce, “Is There Contents in Empty Heads?”, *Proceedings of the 13th International Conference on Computational Linguistics (COLING-13)*, Helsinki, Finland, pp. 138–43, 1990.
- [25] Hindle, D., “Noun Classification from Predicate-Argument Structures”, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, pp. 268–75, 1990.
- [26] Justeson, J. and S. Katz, “Redefining Antonymy: The Textual Structure of a Semantic Relation”, *Using Corpora: Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Oxford, UK, pp. 138–53, 1991.

- [27] Levin, B., “Building a Lexicon: the Contribution of Linguistics”, *International Journal of Lexicography*, 4(3), 205–26, 1991.
- [28] Levin, B., *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, 1993.
- [29] Miller, G. (ed.), Special Issue on WORDNET, *International Journal of Lexicography*, 3(4), 1990.
- [30] Neff, N. and B. Boguraev, “Dictionaries, Dictionary Grammars and Dictionary Entry Parsing”, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, pp. 91–101, 1989.
- [31] Neff, M. and M. McCord, “Acquiring Lexical Data from Machine-Readable Dictionary Resources for Machine Translation”, *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, Austin, Texas, pp. 87–92, 1990.
- [32] Pustejovsky, J., “Towards a Generative Lexicon”, *Computational Linguistics*, 17(4), 1991.
- [33] Pustejovsky, J. and S. Bergler (eds.), *Lexical Semantics and Knowledge Representation*, Springer-Verlag, Berlin, 1992.
- [34] Pustejovsky, J. and B. Boguraev, “Lexical Knowledge Representation and Natural Language Processing”, *Artificial Intelligence*, 63, pp. 193–223, 1993.
- [35] Wilks, Y., D. Fass, C-M. Guo, J. McDonald, T. Plate and B. Slator, “Providing Machine Tractable Dictionary Tools”, in J. Pustejovsky (Ed.), *Semantics and the Lexicon*, Kluwer Academic Publishers, Dordrecht and Boston, 1993.
- [36] Zernik, U. (ed.), *Lexicon Acquisition: Using On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

# Research Toward the Development of a Lexical Knowledge Base for Natural Language Processing\*

Robert A. Amsler  
Bell Communications Research  
*e-mail: amsler@bellcore.com*

## Abstract

This paper documents research toward building a complete lexicon containing all the words found in general newspaper text. It is intended to provide the reader with an understanding of the inherent limitations of existing vocabulary collection methods and the need for greater attention to multi-word phrases as the building blocks of text. Additionally, while traditional reference books define many proper nouns, they appear to be very limited in their coverage of the new proper nouns appearing daily in newspapers. Proper nouns appear to require a grammar and lexicon of components much the way general parsing of text requires syntactic rules and a lexicon of common nouns.

## 1 History

The concept of a Lexical Knowledge Base (LKB) was originated by Amsler and Walker at SRI International during the 1981-1982 period in response to the need for a new approach to natural language processing that could overcome the limitations of traditional computational linguistic analysis of text. It grew out of a number of pessimistic observations of existing parsing systems which indicated that (a) they could not parse and understand text for which they lacked complete lexicon, (b) that manual entry of lexicon of the size needed for unrestricted text analysis in such parsing systems was beyond the capabilities of even well-funded research projects, (c) that manual entry of lexical information duplicating that contained in existing machine-readable dictionaries was a senseless waste of expensive human labor, and (d) that while traditional parsing was required to completely understand text, such a depth of understanding was not required unless one already knew the text was about the topic one wanted to understand and it appeared that some intermediate natural language processing techniques could be developed that would provide this level of "lexical" understanding of text and associated derivable text properties.

But, while less sophistication was needed than that employed in full-scale computational linguistic solutions to the parsing problem, more sophistication was needed than

---

\*This paper is an expanded version of a paper that appeared in the Proceedings of the 1989 SIGIR Conference, Cambridge, MA, 25-28 June 1989.

was currently being provided by information science techniques that had not advanced beyond those developed in the early 1950s by pioneers such as Salton. Information science had early recognized the computer's ability to accurately record and display textual information just as well as numbers. Two early half-truths about the lexicon made text storage and retrieval a technique with hidden fatal flaws.

## 1.1 Two Half-Truths of Information Science

**The first half-truth was that a word was a contiguous sequence of alphabetic characters.**

This seems obvious, but ignored an important classes of words such as open nominal compounds (e.g. ‘ice cream’ or ‘bottle washer’) and phrasal verbs (e.g. *married off* in the sentence, “He married his youngest daughter off.” vs. *married* in the sentence, “He committed incest and married his youngest daughter.”). Idioms were another problem, and a ‘horse of a different color’ caused false retrievals because its content words had nothing much to do with *horses* or *colors*. These difficulties were capable of partial resolution for retrieval of the multi-word lexical items (For instance, if you specified ‘ice cream’ or ‘married\*off’ (assuming \* meant ‘arbitrary sequence of intervening words’) as the desired sequence, you could avoid getting ‘ice’ without ‘cream’ and ‘married’ without ‘off’; but retrieval of the the non-compounded forms of each, (retrieving ‘ice’ when it would not be followed by ‘cream’) wasn’t possible unless one explicitly noted ‘ice’ NOT followed by ‘cream’; but then there was ‘ice milk’ and ‘ice skates’ etc. Specifying all of the NOT’s would take a long time.

**The second half-truth was that computers could retrieve information about a concept by retrieving occurrences of a word form which represented that concept.**

It was obviously true that this worked in many cases. Some word forms are uniquely associated with one and only one concept, and for these word forms information storage and retrieval worked perfectly. If you wanted information about ‘magnetohydrodynamics’ and typed that word into a computer, you got information about that concept back. However, if you wanted information about ‘the production of fiber by plants’ and asked for documents containing ‘fiber’, ‘plants’ and ‘production,’ you might easily find that ‘fiber optic production plants’ would appear. Word forms did not always represent unique concepts.

Related lexical problems were created by many concepts having several word forms which represent them in text. ‘Numismatics’ and ‘coin collecting’; ‘aardvarks’ and ‘ant-eaters’; ‘Strategic Defense Initiative’ and ‘Star Wars’. In addition to this synonymy problem, there was a problem caused by hyponyms. A hyponym is a word which is more specific than a given word; it represents a specific instance of which some other word is the generic concept. In artificial intelligence terminology, it is in an ‘ISA’ (or ‘is a’) relationship to another concept. Thus, a ‘cat’ ISA ‘mammal’; a ‘jeep’ ISA ‘vehicle’; a ‘elm’ ISA ‘tree’, a ‘physicist’ ISA ‘scientist’ and ‘NEW JERSEY BELL’ ISA ‘BOC’. If one asked for information on the employment of scientists by BOCs one would anticipate finding out about physicists employed by NEW JERSEY BELL—but unless one explicitly expanded the choice of word forms to include all the BOCs and

all the types of scientists, this wouldn't happen in contemporary information retrieval systems.

## 1.2 A Cure

This plague of lexical problems, and several others which appear when one endeavors to deal with text, had one common cause. *The systems attempting to process language material lacked a complete lexicon of the language they were attempting to manipulate intelligently and had no rules for understanding how to recognize these lexical concepts when they appeared in text.* This was the classic mistake of information science and one for which it to this day continues to attempt to find a statistical solution.

Computational linguistics has already made considerable progress on representation of lexical concepts. Even as early as 1968, Ross Quillian (Quillian, 1968) had represented information from a dictionary in a form that permitted a computer to describe the meanings of concepts by following labeled directed arcs in a semantic network. During the 1970s the technology of semantic networks grew in the hands of practitioners such as Bob Simmons (Simmons, 1973) and Stuart Shapiro (Shapiro, 1979) while parsing technology evolved. By the 1980s new insights into the structure and use of machine-readable dictionaries emerged in the dissertations of Amsler (Amsler, 1980) and Michiels (Michiels, 1981), and John Sowa (Sowa, 1984) cleaned up the description of semantic networks by restoring them to working order on a par with logical formalisms. Additionally, newer forms of grammar such as Lexical Functional Grammar (Bresnan, 1983), and Categorial Grammar (Wittenburg, 1986), were showing how the specific problems of parsers could be relegated to solutions at the hands of specific lexical entries, rather than requiring omniscient grammatical rules that adequately treated all concepts. All these events seemed to augur that the time was ripe for a new assault on the lexicon.

Unlike the totality of sentence forms, the totality of lexical concepts ought to be within our hands quantitatively and the problem of providing a qualitatively adequate representation of every lexical concept and its relationship to the others was addressable within the framework of ambitious knowledge representation software such as that being developed by Intellicorp (KEE), the Carnegie Group, and the ISI/BBN group (NIKL, Robins, 1986).

## 1.3 A Plan of Action

The task of acquiring the lexicon had two logical starting points. These were to start (a) with machine-readable dictionaries and (b) with raw text. A type of parallel development of both lexical resources and text sources was envisioned as forming the two columns of a new arch that could eventually be built to span their respective lexical contents.

Several machine-readable dictionaries existed and were available by 1980. Three of the most accessible were the *Merriam-Webster Pocket Dictionary* (MPD), its larger sibling, the *Merriam-Webster Seventh Collegiate* (W7) and the *Longman Dictionary of Contemporary English* (LDOCE). The Merriam-Webster and Longman dictionaries offered different capabilities as repositories of data about lexical concepts. The MPD and W7 provided a mature collection of definitions, and the family resemblance of the smaller MPD to the W7 and the W7 to the definitive American English dictionary,

the unabridged *Merriam-Webster Third International* (W3I) provided the ability to find out more about definitions in any of the smaller books by consulting its “big brother” when the need arose. Additionally, the MPD and W7 were the result of an extensive organization effort by a whole series of computational lexicologists who had refined its format to a very easily computed structural description (Reichert, Olney and Paris, 1969; Sherman, 1974b; Amsler and White, 1979; Peterson, 1982b; and Peterson, 1987).

The LDOCE, while very new, offered something relatively rare in dictionaries, a series of syntactic and semantic codes for the meanings of its words. These codes were a fascinating repository of raw linguistic “ore” from which the possibility of additional “finds” could be made.

On the text side, the easiest sources of text were the newswires. In a pioneering step, John McCarthy had acquired a line to the AP and NYTimes news wires at Stanford University in the middle 1970s. SRI acquired its own feed to the NYTimes newswire at the request of Amsler and Walker in 1982. This source of raw text was immediately used to construct an 8 million word corpus of some three months worth of news stories, and to perform isolated word frequency counting on the data so obtained<sup>1</sup>.

Both sides of the arch had presumably been built to such a height that the first comparison of their contents was possible.

## 2 Comparing Lexical Sources with Lexical Resources

The initial comparison was carried out at SRI International in 1983. Its results were surprising and disappointing, as reported in Walker and Amsler (1986):

Of the 119,630 different word forms present in the *NYTNS* sample and in the *W7*, 27,837 (23%) occurred in both; 42,695 (36%) occurred only in the *W7*; and 48,828 (41%) occurred only in the *NYTNS*. The fact that almost two-thirds of the words in the dictionary (actually 61%) did not appear in the text is not surprising; dictionaries contain many words that are not in common use. That almost two-thirds of the words in the text were not in the dictionary (actually 64%) is more problematic.

The following list presents, in different type fonts, a small portion of the total unique form list of 120,000 types of the lexicon that occur (1) both in the *Webster's Seventh New Collegiate Dictionary* *W7* and in three months of text from the *New York Times News Service* (*NYTNS*) [represented in **underlined boldface**]; (2) only in the *W7* [represented in normal type]; and (3) only in the *NYTNS* [represented in *italic type*]. The form of words presented here follows the specifications given in (Walker and Amsler, 1986) and specifically lacks both preceding and trailing punctuation, hence abbreviations do not appear with their trailing periods although such information has been preserved in the actual database.

A preliminary analysis of a sample of the *NYTNS* forms that were not in the *W7* reveals the following breakdown (expressing the values in fractional form is intended

---

<sup>1</sup>It was not yet realized at this point, how critically important open compounds were to the accurate counting of lexical concepts in text.

36%	23%	41%
Dictionary	both	Newswire

Figure 1: Relative Percentages and Categories of Lexicon in 8.35 million word New York Times Newswire Corpus

to show their approximate character): one-fourth were inflected forms; one-fourth were proper nouns; one-sixth were hyphenated forms; one-twelfth were misspellings; and one-fourth were not yet resolved, but some are likely to be new words occurring since the dictionary was published.

25%	25%	14%	8%	25%
Proper Nouns	inflected forms	hyphenated	misspell	unknown

This suggested a serious incompatibility between newswire text and this dictionary. The incompatibility appeared to depend critically on acquiring additional proper noun lexicon as well as handling inflected forms (morphology) and hyphenation (distinctions between hyphenated (e.g., “data-base”) and closed (e.g., “database”) compounds), and spelling correction.

Additionally, a new problem for consideration was that isolated words, especially when extracted from proper nouns, were not representative of the lexical coverage of a text. It made no sense to consider “New” as a word when the text had actually contained “New Zealand”. Thus any true assessment of the conceptual coverage of a lexicon relative to the contents of a text corpus would really depend upon the coverage of open compounds rather than just isolated words. The open compounds in a machine-readable dictionary were readily found since they would appear as main entries, however the open compounds contained in a text corpus were not obvious. New programs were needed to attempt to find open compounds in text.

**2.1 Finding Open Compounds in Text** The task of finding the open compounds in text which ought to have lexical entries is a very difficult one. Even the relatively easier task of finding open proper noun compounds in text, in which one can rely upon capitalization of the key elements of the compound, is not without problems. Proper nouns can contain intermediate words such as ‘and’, ‘the’, ‘of’, etc., which also can appear between two or more distinct open compounds. Thus “United States of America” and “American Telephone and Telegraph Company” are single proper nouns, whereas “Mayor Koch of New York City”, and “United States and Canada” are each composed of two proper

..., g, g&w, g-6, g-g-c, g-man, g-r-i, g-rated, g-string, g-t-c, g.a, g.b, g.d, g.e, g.h, g.i, g.k, g.m, g.o.p, g.p, g.v, ga, gaafar, gab, gaba, gabardine, gabathuler, gabbed, gabber, gabbi, gabbing, gabble, gabbler, gabbro, gabbroic, gabbroid, gabby, gabe, gabelle, gabelli, gaberdine, gaberlunzie, gabfest, gabino, gabion, gable, gabled, gabler, gables, gabo, gabon, gaboon, gabor, gabriel, gabriela, gabriele, gabriella, gaby, gacy, gad, gadabout, gadacz, gadarene, gadded, gader, gadding, gadfly, gadget, gadgeeteer, gadgetry, gadgets, gadgety, gadite, gadoid, gadolinite, gadolinium, gadroon, gadrooning, gadsden, gadusek, gadwall, gae, gaea, gael, gaelic, gaels, gætan, gaf, gaff, gaff-top sail, gaffe, gaffer, gaffes, gaffing, gaffney, gag, gaga, gagarin, gagarin's, gage, gagged, gagger, gaggle, gagman, gags, gagster, gahnite, gai, gaiety, gail, gailani, gailey, gaillardia, [gaily], gain, gaine, gained, gainer, gainers, gaines, gainesville, gainey, gainful, gainfully, gainfulness, gaingiving, gaining, gainless, gainlessness, gainly, gains, gainsay, gainsayer, gainsborough, airy, gait, gaited, gaiter, gaithersburg, gaitley, gajowniczek, gal, gal's, gala, galact-, galactic, galacto-, galactopoiesis, galactopoietic, galactose, galactoside, galah, galahad, galambos, galanos, galantine, galapagos, galatea, galavant, galax, galaxies, galaxy, galbanum, galbraith, galbraith's, galbreath, galda, galdogob, gale, galea, galeate, galeiform, galella, galen, galena, galenic, galenical, galenism, galeo, galerie, gales, galesburg, gali, galil, galilean, galilee, galileo, galileo's, galimatas, galimore, galina, galingale, galiot, galipot, gall, galla, gallagher, gallant, gallantly, gallantry, gallas, gallaway, gallbladder, galleass, gallegos, gallein, gallen, gallen's, gallen-kallela, galleon, galleria, galleried, galleries, gallery, gallery's, galleta, galley, galley-west, galleys, gallfly, galliani, galliano, galliard, gallic, gallican, gallicanism, galicism, gallicize, gallicolous, gallienne, galligaskins, gallimaufry, gallinaceous, galling, gallinipper, gallinule, gallios, galliot, gallipot, gallium, gallivant, gallivorous, gallix, gallnut, gallo, gallo's, gallois, gallon, gallonage, gallons, galloon, gallop, galopade, galloped, galloper, gallophile, galloping, gallops, galloway, gallowglass, gallows, gallstone, gallstones, gallup, gallus, gallused, gally, galoot, galop, galore, galosh, galoshed, galoshes, gals, galston, galt, galtieri, galtieri's, galumph, galvan, galvani, galvanic, galvanically, galvanism, galvanization, galvanize, galvanized, galvanizer, galvanizing, galvano, galvano-, galvanometer, galvanometric, galvanoscope, galveas, galveston, galveston's, galvin, galway, galyak, gam, gam-, ...

Figure 2: Comparison of the New York Times and Dictionary Entries

nouns. Furthermore, contractions such as “North and South Carolina” can complicate any assessment of lexical coverage in a corpus.

Added to this confusion over contiguous words, there are two additional problems provided by (1) the inability to distinguish between abbreviation periods and periods ending sentences (or, in the worst case, a single period serving both functions) and (2) the requirement that the initial word of sentences always be capitalized. Thus, any single word can appear to be a proper noun if it can start a sentence, and any sentential period can be misinterpreted as an abbreviation period between two words which are part of one open proper noun compound.

The scene now changes to Bellcore as Amsler and Walker relocated to the East coast to resume this research. At Bellcore, additional reference works were acquired to augment the needed missing lexicon. One work which seemed to be essential and which was acquired was a 1985 edition of *The World Almanac and Book of Facts* (WA85) in machine-readable form. This was thought to be an ideal source of proper nouns since the WA85 was a highly compact book containing highly competitively selected proper nouns with known information salience. That is, while not giving exhaustive lists of all

biographic, geographic, or other proper noun classes, the WA85 selected those proper nouns from each class that were most likely to be known. This seemed to provide a good basis for understanding something as general as a newswire.

In order to compare proper noun coverage of WA85 and the NYTimes, it was necessary to perfect a program to extract proper noun compounds and isolated proper nouns from text, effectively working without a lexicon except for some list of “stopwords”.

**2.2 CAPS.SNO** Taking into account the difficulties of finding proper nouns, a program with many of the desirable heuristics was written in SNOBOL4<sup>2</sup> The program operates as follows:

1. Line-delimited upper and lower case text is read in and searched for successive *sentences*. A *sentence* is heuristically described as an arbitrary sequence of text terminating with two non-blank characters followed by a period which is itself followed by either two single quotes, (""), (as universally used in newswire text to indicate quoted text) or a blank.

Thus all of the following three would be considered *sentences*,

(This is a sentence. “This is a second sentence.” This a third sentence.)

but, the following quoted terms would be seen as being inside the larger *sentence* shown,

(“Star Wars,” also known as the “Strategic Defense Initiative,” would require an enormous expenditure on space weapons research.)

2. Within a *sentence*, a *proper noun* is heuristically defined as one of:

- an arbitrarily long sequence of ‘capitalized words’ separated by single blanks from each other, or
- a sequence of two ‘capitalized words’ separated by a sequence of single blank-separated acceptable non-capitalized ‘joining words’ (which currently are: ‘of,’ ‘the,’ ‘on,’ ‘a,’ ‘an,’ ‘to,’ ‘for,’ and ‘and’), or
- a single ‘capitalized word’ followed by a terminating punctuation symbol (which currently includes blank, comma, semicolon, colon, double-quotes, exclamation mark, question-mark, period, equals-sign, or a closing angle-bracket, curly-bracket, square-bracket or parenthesis), or
- a sequence of one or more of any of the above in any order.

---

<sup>2</sup>SNOBOL4 is an extremely useful language for dealing most large text processing tasks. It is interpretive, rather than compiled, permitting run-time dynamic memory allocation, and has built-in functions for most needed text manipulation tasks. It is also highly portable and standardized such that SNOBOL4 programs written on machines as different as TOPS-10/20-based DEC-10/20's, UNIX-based VAXes, and MS-DOS-based IBM PCs can all execute the same programs with minor I/O modifications.

3. A ‘capitalized word’ is any sequence of upper or lowercase letters or hyphens and periods which begins with one of the following:

- a single capitalized letter, or
- a single capitalized letter preceded by a double quote, (“”), (indicating the capitalized word started a quotation in newswire text), or a sequence of ‘initials’ (single capitalized letters followed with periods)

This program was also modified to run on the *World Almanac and Book of Facts* 1985 (WA85) data, which contains material in tabular format, rather than as sentences. The entire text of the WA85 and one month’s worth of the NYTimes was run through the proper-noun extraction program. However, once again, the intersection of the resulting data was disappointing. The WA85 contained less than 10% of the proper nouns in the newswire text. Figure 3 shows a sample of the proper noun coverage. Proper nouns in *italic type* are only in the newswire, those in normal Roman type are only in the *World Almanac* and those in **underlined boldface** are in both the *World Almanac* and the *New York Times*. Processing more newswire data would of course increase the size of the intersection, but would not change the number of the proper nouns from the first month which did not appear in WA85.

Clearly something was going on here that needed explanation. How could human readers understand text if they only knew as little as 10% of the actual proper-noun lexical entries in the text?

Examining the data more closely it became clear that there was a more complicated problem here than a simple mismatch of lexical coverage between WA85 and the NYTimes. While indeed there were many proper nouns in the NYTimes that had no counterparts in the WA85, many of the proper nouns were also ‘near misses’ of each other. One source abbreviates when the other does not. One provides the full name when the other provides a shorter form of the same name. Unlike the situation with common nouns, these differences were not just those of known linguistic morphology. That is, the proper nouns obeyed what appeared to be grammatical rules with which few linguists have dealt.<sup>3</sup>

There was no easy way to detect a composite proper noun, that is, one composed of two or more proper nouns concatenated together, unless one actually has the separate component pieces in their exact component forms. And composite proper nouns appeared to have both parseable and non-parseable (i.e., lexical) components. Thus, “Grand Forks Energy Research Center” might be taken as a composite proper noun with “Grand Forks” being a proper noun and “Energy Research Center” being a common noun specialized for this instance. That is presumably how human readers would be able to accept this proper noun when they encountered it for the first time. Thus by accepting “Grand Forks” as a placename, and “Energy Research Center” as a meaningful content phrase, being a ‘center’ at which ‘research’ is carried out on ‘energy’, the reader can accept this phrase without previously having had a lexical entry explaining its meaning. The situation is far different for something such as “Grand Funk Railroad” which without a

---

<sup>3</sup>Among those who have noticed this problem are John Carroll (Carroll, 1985), and a small group of linguists who study an area termed ‘lexicalist morphology’ (Botha, 1984).

... *Graham Greene and John Le Carre*, *Graham Greene's*, *Graham Greene*, Graham H. Rights, **Graham Hill**, *Graham Kemp of Input Inc.*, *Graham Main*, *Graham Nash*, *Graham Payn and Sheridan Morely*, *Graham Payn and Sheridan Morley*, *Graham T. Allison*, *Graham Whitehead*, *Graham's Land*, **Graig Nettles**, *Grain Centers*, *Grain Millers*, *GRAIN PREPARATION*, *Grain Receipts*, *Grain Storage Capacity*, *Grain Store Capacity*, *Gram Parsons'*, *Grambling State Univ.*, *Grambling State*, *Grambling University*, *Gramercy Arts Theater*, *Gramercy Park South*, **Gramercy Park**, *Gramm and Conable*, *Grammy Awards*, *Grammy Award*, *Gran Canaria*, *Gran Chaco*, *Gran Frau*, *Gran Fury*, *Gran Paradiso*, *Gran Sasso Tunnel*, *Gran Via*, *Gran Vina Sol*, *Granat and Mickelson*, *Granby St.*, *Grand Admiral Karl Doenitz*, *Grand Alliance*, *Grand and Dodier*, *Grand Anse Beach*, *Grand Army of the Republic*, *Grand Army Plaza*, *Grand Ave.*, *Grand Avenue Mall*, *Grand Avenue*, *Grand Ave*, *Grand Bahama Island*, *Grand Ballroom of the Conrad Hilton Hotel*, *Grand Banks*, *Grand Blanc*, *Grand Bourg Cemetery*, *Grand Bourg's*, *Grand Bourg*, *Grand Canary*, *Grand Canyon of the Colorado*, *Grand Canyon State*, *Grand Canyon Suite*, **Grand Canyon**, *Grand Cayman Islands*, *Grand Cayman Island*, *Grand Cayman*, *Grand Central Sta.*, *Grand Central Station*, **Grand Central Terminal**, *Grand Central Ter*, *Grand Central*, *Grand Circle Travel of New York*, *Grand Circle Travel*, *Grand Circle*, *Grand Coalition*, *Grand Combin*, *Grand Comoro*, *Grand Coulee Dam*, **Grand Coulee**, *Grand Coulee*, *Grand Duchy of Luxembourg*, *Grand Duchy of Warsaw*, *Grand Duke and Duchess*, *Grand Duke Jean*, *Grand Duke Michael*, *Grand Duke of Lituania*, *Grand Dukes of Muscovy*, *Grand Dukes of Vladimir*, *Grand Dukes*, *Grand Dutchesses Olga*, *Grand Encampment*, *Grand Forks Energy Research Center*, *Grand Forks*, *Grand Funk Railroad*, *Grand Guignol*, *Grand Gulf*, *Grand Hall of Columns*, *Grand Haven*, *Grand Hotel on Mackinac Island*, *Grand Hotel*, *Grand Hyatt*, *Grand Island*, *Grand Islan*, *Grand Isle*, *Grand Junction and Montrose*, **Grand Junction**, *Grand Junct*, *GRAND JURY*, **Grand Jury**, *Grand Ledge*, *Grand Lodges*, *Grand Marai*, *Grand Marnier*, *Grand Master Flash*, *Grand Master Nebula Award*, *Grand Mere*, *Grand Mesa Forest*, **Grand Mosque**, *Grand Moulins*, *Grand National Assembly*", *Grand National Champion-NASCAR*, *Grand National Champions*, *Grand National*, *Grand Ol' Opry*, *Grand Old Flag*, *Grand Old Opry*, *Grand Old Party*, **Grand Ole Opry**, *Grand Opera*, *Grand Patron*, *Grand Place*, *Grand Portage*, *Grand Prairie*, *Grand Prix for Formula*, *Grand Prix Formula*, **Grand Prix**, *Grand Prize*, *Grand Rapids Baptist Coll*, *Grand Rapids Jr.*, *Grand Rapids Press*, *Grand Rapids Symphony*, **Grand Rapids**, *GRAND RAPIDS*, *Grand Rapid*, *Grand Slams*, *Grand Slam*, *Grand Terrace*, **Grand Teton National Park**,  
...

Figure 3: Comparison of Proper Noun Lexicon in World Almanac and New York Times

lexical entry would be assumed to be the name of a railroad, possibly owned by someone or located near someplace named ‘Funk’, rather than a rock group.

Thus, what was discovered here was that proper nouns have grammatical structure. They have alternate forms based upon contractions and occasionally even rearrangements (e.g., “Speaker of the House” vs. “House Speaker”)<sup>4</sup>. They can be understood in part from stored lexical entries for proper nouns, and in part from allowing what appear to be proper nouns to assume their common noun meaning. This makes recognition of proper nouns a very hard task requiring effectively a small-scale parsing operation to take place in order to understand them. I have expressed this by saying that proper nouns are not quite ‘lexical’ in nature, since enumeration of all proper nouns prior to parsing appears to be impractical<sup>5</sup> if not actually impossible<sup>6</sup>. That is, one would have to enumerate all forms of the proper nouns if one doesn’t admit to needing grammatical mechanisms for their representation.

**2.3 Role of Text Sources in an LKB** If one cannot “find” ready-made lexicon in dictionaries that will match the lexical concepts appearing in text, then a successful strategy for completing a lexical knowledge base will be to build adequate compositional lexical entries that can serve the needs of a suitable lexical parser. What then, one may well ask, is the justification for collecting both text and machine-readable dictionaries. The answer is that both continue to be valuable assets for building a lexical knowledge base. In the first place, one needs a complete compositional lexicon such that the parser can operate. In the above example of “Grand Funk Railroad” and ‘Grand Forks Energy Research Center’, the lexicon must already contain the first term in its entirety, and the component parts of the latter expression, i.e. “Grand Forks”, and suitable entries for ‘energy’, ‘research’ and ‘center’. However, the compositional rules for such compounds as “Grand Forks Energy Research Center” will have to be written. To write such rules will require understanding all the compound types in which the word ‘center’ will appear.

Thus, while a “recreation center” is a “a building or part of a building used for recreation,” a “garden center” is not “a building or part of a building used for a garden,” but instead a “a store or establishment devoted to gardening, carrying supplies, materials, tools, and books as well as offering guidance and advice.” To write a correct lexical entry for the concept of “center” as it is used in text, one needs to have a virtually complete collection of examples of different uses. Figure 4 shows a small sample of the 560 open-compound proper-noun uses of ‘Center’ from the NYTimes corpus.

What immediately becomes apparent in examining this list is that there are many classes of centers which use the same proper-noun compound, such as “Medical Center”. Some such centers are in dictionaries—though they are hard to find because dictionaries do not cross-index them and one has to guess the initial words under which they are presented. This in itself is an excellent argument for the need for machine-readable

---

<sup>4</sup>A further exposition of this trait of proper nouns was presented at the *Third Workshop on Theoretical Issues in Natural Language Understanding*, held at New Mexico State University in Las Cruces, Jan 7-9, 1987; see Amsler (1987a).

<sup>5</sup>Many are rearrangements or contractions of other forms, thus wasting perhaps five times as much space per proper noun for a purely enumerative listing of all forms possible.

<sup>6</sup>If humans actually read text without having a complete proper noun lexicon, then assuming one can build such a lexicon for a computer may be untrue and it may be that parsing of proper nouns is required for an acceptable level of understanding.

"Epcot Center"	Center	on PBS
"Live From Lincoln Center"	Center	Leaves
"Luckies Use Only Abbot Cruise Center	Center	
Afghan Information Center	Center	
Agency for International ...		
Agriculture Department's Northern Regional Research Center	Center	
Air France Brochure Distribution Center	Center	
Akron Center	Center	for Reproductive Health
Alexian Brothers Medical Center	Center	
Allen Park Youth Center	Center	
Ambrose Housing Center	Center	
American Medical Association's Center	Center	for Health Policy Research
American Productivity Center	Center	
Ames Research Center	Center	
Anaheim Convention Center	Center	
Ariel Merari of Tel Aviv University's Center	Center	for Strategic Studies
Armed Forces Medical Center	Center	
Asia Foundation's Translation Service Center	Center	
Athletic and Convocation Center	Center	
Atlanta Memorial Arts Center	Center	
Atlantic City Convention Center	Center	
Atmospheric Sciences Center	Center	
Atrium of the Kennedy Center	Center	of the State
Baltimore Civic Center	Center	University...
Baltimore's Civic Center	Center	
Barbican Center	Center	
Bell Phone Centers	Centers	
Berkshire Music Center	Center	
Bernard Horwitz Jewish Community Center	Center	
Biosexual Psychohormonal Clinic of Johns Hopkins Medical Center	Center	
Border Studies Center	Center	for the Arts
Boston Center	Center	
Boston University Medical Center	Center	for Adaptive Systems
Boston University's Center	Center	for Design
Boston's Center	Center	
Briarbrook Common Shopping Center	Center	
Briarbrook Commons Shopping Center	Center	
Britain's Driver and Vehicle Licensing Center	Center	
Brookings Institution Foreign Policy Studies Center	Center	
Studies Center	Center	
Brooklyn Detention Center	Center	

Figure 4: Sample of Proper Nouns Uses of 'Center' in Sept.-Nov. '82 NYTimes

dictionaries, which permit access by any component of a lexical entry, rather than only its initial word.

Figure 5 shows a summary of the open compounds containing ‘Center’ from the NYTimes corpus together with hypotheses as to their domains of use.

Democratic center [Politics]	health center [Medical]
“Center for ...”	health science center [Medical]
“Center for the Arts” [Arts]	health sciences center [Medical]
“center for ... research”	heart and lung center [Medical]
“center for ... studies”	home care center [Medical]
“center for research”	hospital center [Medical]
arts center [Arts]	housing center
athletic center [Health]	information center
banking center [Business]	law center
center court [Tennis]	law enforcement center
center party [Politics]	linguistic center
children’s center	medical center [Medical]
civic center [Civic]	mental health center [Medical]
communications center [Industry]	music center [Arts]
community center [Civic]	nutrition center [Medical]
conference center [Civic]	performing arts center [Arts]
convention center [Civic]	physical fitness center [Health]
correctional center [Criminology]	poison center [Medical]
cruise center [Tourism]	policy center
cultural center [Arts]	production center [Business]
design center [Arts]	refugee center
detention center [Criminology]	research center
distribution center [Industry]	shopping center [Business]
documentation center	space center [NASA]
drama center [Arts]	space flight center [NASA]
education center [Education]	sports center [Sports]
entertainment center [Business]	tennis center [Sports]
exhibition center [Civic]	theater center [Arts]
exposition center [Civic]	trade center [Business]
fitness center [Health]	training center
garment center [Business]	transplant center [Medical]
geriatric center [Medical]	welcome center [Civic]
graduate center [Education]	youth center [Community]

Figure 5: Summary of Open Compounds containing ‘center’ in NYTimes corpus

An important property exhibited here is that although these lexical items are proper nouns, we are able to find the “meanings” for some of their components as if they were just common nouns. This is significant in that a proper noun doesn’t have to be semantically meaningful in its components—it can be, as “Grand Funk Railroad”, non-decomposable, or it can in fact freeze any linguistic structure in its title, as “Graham’s Land,” which uses a possessive as part of a proper noun. Thus, while there is no guarantee that a proper noun can be semantically decomposed—a large number of proper nouns are understood by first being converted into common nouns and then semantically parsed for the meaning of the common noun phrase. “Graham’s Land” becomes “Graham’s land” which means “land belonging to Graham.” We do this without thinking about it

unless such a process is blocked because of the existence of specific knowledge in our memory that the term cannot be decomposed. "New England" is no longer parsed as "new England," that is, "a new version of England."

### 3 Open-Compound Common Nouns

This provides a convenient bridge to a discussion of the still harder problem of open-compound common nouns in text. Whereas proper nouns can be recognized by their capitalized word components, open-compound common nouns (e.g. "ice cream sundae", etc.) pose a harder task for lexical recognition. A new program, 'phrases.sno,' was written to look for open-compound common-noun phrases in text.<sup>7</sup> The output of this program consists of two sets of data, multi-word open-compound phrase candidates and (optionally) isolated non-function words. In effect the program attacks one of the serious problems of automatic indexing of text by attempting to find multi-word objects to be indexed rather than treating every word as an isolated occurrence and thereby splitting apart true compounds. However, the program is not reliable enough for unedited use. This is why the phrases produced are termed 'candidates' rather than actual phrases.

**3.1. PHRASES.SNO** Phrases.sno extracts either just multi-word lexical phrase candidates or both multi-word and isolated single word lexical candidates. This is an option offered at the start of the program.

Phrases.sno uses a different algorithm than caps.sno. Whereas caps.sno first selects whole sentences in which to find a proper noun phrase, phrases.sno builds phrases word by word. The essential element of phrases.sno is the *symbol*, which is any sequence of non-blank characters. Phrases.sno has two fundamental states. In the first state, it is seeking an acceptable initial content symbol in the text which might start a phrase. In the second state, it has found one or more acceptable initial symbols and is seeking additional consecutive acceptable symbols.

1. Phrases.sno begins by reading in lines of data until it finds a non-blank-line. It then removes any punctuation, numbers, blanks or tabs which begin the line and checks again to determine that the line is still non-blank. The punctuation removed includes any of:

. ' ! ' ' # \$ % & ( ) \* = \_ : - ' @ + ; < , > ? / [ ] ^ ~ { } |

2. Next, an initial symbol, SYMBOL1, is removed from the front of the line and any following blanks and tabs are discarded. SYMBOL1 is checked to determine that it does not consist solely of numbers and punctuation, although it can contain these at this stage.
3. Passing this test, SYMBOL1 it is tested to see if it consists of two characters terminated with a span of punctuation. If so, this SYMBOL1 is considered to be

---

<sup>7</sup>This program was used by Lynn Streeter, and Susan Dumais for other phrase extraction experiments at Bellcore. It has also been implemented in C by Karen Lochbaum.

an isolated content word and output as such (if this option has been selected for output).

4. The SYMBOL1 is now checked to see if it is either of the letters 'A' or 'I' followed by a period. 'A' and 'I' are the two one-letter function words of English, but followed by a period they are also capable of being a person's initial, as such should not be discarded. If SYMBOL1 is either of these two, it is retained and processing passes to the second state of the program—hunting for a span of two or more consecutive content words.
5. Granted SYMBOL1 is not 'A.' or 'I.', then it is reduced to lower-case letters and checked to see if it is a known 'function' word. The list of function words is enumerated at the start of the program and as the code is interpretive, can be easily edited by the user to add/delete other words. The program by default considers the following to be function words:

a, about, according, after, against, ago, all, also, although, among, am, an, and, another, any, are, around, as, at, back, be, because, become, been, before, being, between, both, but, by, called, came, can, come, could, did, do, don't, down, during, each, either, even, every, few, for, former, from, get, go, going, got, had, has, have, he, her, here, him, his, how, i, if, in, into, is, it, it's, its, just, know, last, less, like, little, made, make, many, may, me, might, miss, more, most, much, must, my, never, next, no, not, now, of, on, onto, one, only, or, other, our, out, over, own, put, said, same, say, says, see, several, she, should, since, so, some, still, such, take, than, that, the, their, them, then, there, these, they, this, those, through, to, together, told, too, under, until, up, use, used, very, want, was, way, we, well, were, what, when, where, which, while, who, will, with, within, without, would, you, your.

6. If SYMBOL1 is one of these words, it is discarded and the next word on the current line is examined. If there are no further words on the current line, then the next line of data is read and the program begins to process that line as it did the first.
7. Otherwise, SYMBOL1 is considered to be a good first content word and processing proceeds to find its immediately consecutive symbol, SYMBOL2. SYMBOL2 is extracted using the same pattern used for extracting SYMBOL1 above. If SYMBOL2 also passes the tests applied to SYMBOL1 then a string, SYMBOLIST is created containing both SYMBOL1 and SYMBOL2 separated by a blank.
8. If SYMBOL2 fails the test because it is not a content word, then SYMBOL1 will be output as an isolated content word (if this option was selected). If SYMBOL2 fails because it terminates in punctuation, then SYMBOLIST will be output as a two-word content phrase candidate.
9. Otherwise SYMBOLIST will be retained and the program will search for additional symbols, spanning line boundaries, as long as acceptable additional symbols

are found. Whenever another acceptable symbol cannot be added, the cumulative contents of SYMBOLIST will be output.

**3.2. Common-Noun Phrase-Candidate Validation** To validate the phrases extracted by the program required a different strategy than that employed for proper nouns. The proper noun extraction program was sufficiently accurate that it could simply be applied to a reference work, such as the *World Almanac*, and would extract the proper nouns mentioned in that work for comparison with those proper nouns extracted from newswire text sources. This wouldn't work for the phrases program. It extracted everything that 'could' be a phrase—but had no means to eliminate false phrase candidates.

The strategy selected was to amass all the known open compounds from the main entries in our existing dictionaries. These compounds would then be compared to the phrase candidates extracted by the phrases program. The dictionaries available included a word list from the *Merriam-Webster Second International* (W2I), the *Merriam-Webster Seventh Collegiate* (W7), and the *Collins Dictionary* (CED). The task of extracting open-compound main entries from a machine-readable dictionary is readily performed, and the three sources mentioned yielded a total of 61,201 different open compounds.

The next task, comparing these compounds to the phrase candidates, required some planning. In the first place, the comparison was not likely to be an exact match in most cases—it would most likely match just some part of the phrase candidate, thus if "shopping center" had an entry in the CED dictionary, the phrase "suburban shopping centers" might occur in the phrase candidates. It was desirable that the longest matching sub-expression be identified and displayed with its source dictionaries.

In the second place, there was an enormous number of both phrase candidates (an average of 195,000 per month) and dictionary open compounds (61,201). Some methods of comparison would be very inefficient. For example, comparing open compounds three words in length to phrase candidates only two words long was senseless. Also, one wanted to match the longest open compound from the dictionary before matching any possible shorter compounds (e.g., "United States of America" should be matched before "United States").

Representations of both the phrase candidates and the dictionary compounds which included their lengths were devised and then sorted into the appropriate length orderings for use as single pass batch algorithm data. In a more advanced system it would of course be desirable to use a hash-coding scheme (such as LISP offers) for retrieval of the dictionary data, but at this stage the output from a single pass was all that was desired.<sup>8</sup>

Figures 6 (proper nouns) and 7 (common nouns) show samples of the merged open-compound entries of the three dictionaries (W2I, W7 and CED) in their easily sorted length-encoded representation.

The format of the entries is as follows:

*<word-1> <number> <word-2>...<word-n>. <dictionary-list>*

---

<sup>8</sup>Initially, an attempt to perform the comparisons in regular Franz LISP on the VAX 8650 was undertaken only to discover that there was insufficient memory available to hold the 61,201 open compounds (since a LISP machine wasn't conveniently available and we didn't have the "biglisp" version of Franz LISP installed). A simpler strategy was needed which used older batch-style algorithms to compare the files.

where

*<word-1>* is the initial word of a multi-word open compound,

*<number>* is the number of words in the multi-word open compound,

*<word-2>* through *<word-n>* are the remaining words of the open compound

*<dictionary-list>* are the abbreviated names of the dictionaries in which the open compound appears as an entry

(W7 = *Merriam-Webster Seventh Collegiate*)

(W2I = *Merriam-Webster Second International (word list)*)

(CED = *Collins English Dictionary*)

Graham 2 Land . CED	Grand 2 Coulee . CED
Graian 2 Alps . CED	Grand 2 Falls . CED
Gram's 2 method . CED	Grand 2 Guignol . CED
Grampian 2 Mountains . CED	Grand 2 Lama . CED W7
Grampian 2 Region . CED	Grand 2 Manan . CED
Gran 2 Canaria . CED	Grand 2 Master . CED
Gran 2 Chaco . CED	Grand 2 Mufti . CED
Gran 2 Paradiso . CED	Grand 2 National . CED
Gran 2 chimu . W2I	Grand 2 Pre . CED
Grand 3 Old Party . CED	Grand 2 Prix . CED
Grand 2 Bahama . CED	Grand 2 Rapids . CED
Grand 2 Banks . CED	Grand 2 Remonstrance . CED
Grand 2 Canal . CED	Grand 2 guignol . W2I
Grand 2 Canary . CED	Grandma 2 Moses . CED
Grand 2 Canyon . CED	

Figure 6: Sample of Merged Open-Compound Proper-Noun Lexicon from three Machine-Readable Dictionaries

Figure 8 shows a sample of the output produced by the effort to match common-noun compounds with open-noun compounds in the NYTimes newswire text. Partial matches are represented by a square-bracketed sequence of two or more words within the entire candidate expression. Using this strategy a total of 4,410 phrases were identified within the 366,862 candidate phrases of the corpus. This percentage is still quite small and further consideration of which phrases ought to have lexical entries (such as for proper nouns) and how these can be found is continuing.

Most recently, Yaacov Choueka has developed an alternate algorithm for the identification of phrase candidates, and a study is underway to compare the efficiency of both algorithms. Additionally, the whole question of when and how a phrase should be given a lexical entry is at issue. Lexicographers are clearly much more conservative in creating compound lexical entries than computational linguists, information scientists, or knowledge-based system designers.

The format of the entries is as follows:

*<number1> <number2> <letter(s)> <number3> <phrase candidate>*

graham 2 flour . CED W2I W7	grain 2 thresher . W2I
grain 2 alcohol . CED W2I W7	grain 2 tin . W2I
grain 2 aphid . W2I	grain 2 traveler . W2I
grain 2 beetle . W2I	grain 2 weevil . W2I
grain 2 bill . W2I	grain 2 weigher . W2I
grain 2 binder . W2I	grain-wagon 2 hitch . W2I
grain 2 borer . W2I	grains 3 of paradise . CED
grain 2 broker . W2I	gram 2 atom . CED W2I W7
grain 2 carrier . W2I	gram 2 calorie . CED W2I
grain 2 cleaner . W2I	gram 2 degree . W2I
grain 2 cradle . W2I	gram 2 equivalent . CED W2I
grain 2 crusher . W2I	gram 2 ion . W2I
grain 2 drill . W2I	gram 2 molecule . CED W2I W7
grain 2 elevator . CED W2I	gram's 2 method . W7
grain 2 farm . W2I	gram-atomic 2 weight . CED
grain 2 farmer . W2I	gram-equivalent 2 weight . CED
grain 2 farming . W2I	gram-molecular 2 weight . CED
grain 2 founder . W2I	grama 2 grass . CED
grain 2 glove . W2I	gramicidin 2 D . CED
grain 2 gold . W2I	grammar 2 college . W2I
grain 2 grower . W2I	grammar 2 school . CED W2I W7
grain 2 harvester . W2I	grammatical 2 meaning . CED W7
grain 2 huller . W2I	gran 2 cassa . CED
grain 2 lac . W2I	gran 2 torismo . CED
grain 2 leather . W2I	grana 2 cheese . W2I
grain 2 louse . W2I	granadilla 2 tree . W2I
grain 2 mark . W2I	granary 2 weevil . W2I
grain 2 mash . W2I	grand 2 chain . CED
grain 2 merchant . W2I	grand 2 climacteric . W2I
grain 2 miller . W2I	grand 2 duchess . CED W7
grain 2 moth . W2I	grand 2 duchy . CED W7
grain 2 musk . W2I	grand 2 duke . CED W7
grain 2 oil . W2I	grand 2 final . CED
grain 2 pan . W2I	grand 2 jury . CED W7
grain 2 rust . W2I W7	grand 2 juryman . W2I
grain 2 sack . W2I	grand 2 larceny . CED W7
grain 2 sacker . W2I	grand 2 mal . CED W7
grain 2 sampler . W2I	grand 2 opera . CED W7
grain 2 screen . W2I	grand 2 piano . CED W7
grain 2 screener . W2I	grand 2 seigneur . CED
grain 2 shipper . W2I	grand 2 siecle . CED
grain 2 side . W2I	grand 2 slam . CED W2I W7
grain 2 smut . W2I	grand 2 tour . CED W7
grain 2 soap . W2I	grand 2 vizier . CED
grain 2 sorghum . W2I W7	

Figure 7: Sample of Merged Open-Compound common-noun Lexicon from three Machine-Readable Dictionaries

2 4 H 6 [Grand Guignol . CED] bloodbath starring Dennis Hopper  
 1 1 H 3 [Granny Smith . CED] apples  
 1 1 I 2 [Granny Smith . CED]  
 1 1 I 5 [grade school . CED W7] reading series published  
 1 1 H 8 [grand jury . CED W7] indicted Terp basketball players T. Long  
 1 1 H 5 [grand jury . CED W7] rumors he'd heard  
 2 2 I 5 [grand jury +'s . CED W7] term expires Jan  
 1 1 I 2 [granulated sugar . CED]  
 1 1 H 4 [grape growing . W2I] became overwhelming  
 1 1 H 5 [graphic arts . W7] material group accounted  
 1 2 H 4 [graphic arts . W7] material group  
 1 1 H 4 [graphic arts . W7] material unit  
 1 1 H 2 [graven image +s . CED W2I]  
 1 1 I 4 [graveyard shift . W2I] resent Kennedy  
 1 1 I 3 [graveyard shift . W2I] again  
 1 1 H 3 local [grammar school . CED W2I W7]  
 2 2 H 4 stirring [grass roots . CED W7] opposition  
 1 1 I 4 uncoordinated anti-drug [grab bag . W2I]  
 1 1 H 5 movie's final [grace note . CED W2I W7] suggests  
 1 1 I 4 publicly owned [grain elevator +s . CED W2I]  
 1 1 I 5 five special [grand jury . CED W7] investigations  
 1 1 H 5 Orange County [grand jury . CED W7] met  
 1 1 H 5 Madison Farm Service [grain elevator . CED W2I]  
 1 1 I 5 Moscow-born Tchelistcheff learned [grape growing . W2I]

Figure 8: Sample of phrases.sno Output flagged for Known Open Noun Compounds

where,

<number1> is the number of days on which the <phrase candidate> occurred,  
 <number2> is the number of times the <phrase candidate> occurred,  
 <letter(s)> are a representation of the months during which the <phrase candidate> occurred, with A=January, B=February, ... H=September, I=October, ...  
 <number3> is the length of the <phrase candidate>, and  
 <phrase candidate> is the tagged open compound which phrases.sno extracted and which contains the [,] delimited dictionary main entry and dictionary source acronyms.

## 4 The Future: How to Measure Progress in LKB Work

Measuring progress in a field that still doesn't have any examples of its final output is bound to be hard. Not only is there no agreement as to what a lexical knowledge base should look like, but there probably will never be such a universal agreement. However, despite this, there are some measures against which progress can be judged. A few of these are:

- 1. Lexical Coverage of Unrestricted Text** The issue here is whether one is closing the gap toward having in a lexical database (not necessarily even a knowledge base) all the lexical concepts which appear in some text stream which continues to be produced and against which the actual rate of growth of new lexicon can be measured. There are other topics here, such as how does one decide what a “lexical concept” is, given the debate over proper nouns and other open compounds. An answer is, regardless of whether one believes these forms to be lexical or productive “sentences” in some restricted grammar or words, they do need to be accounted for. Thus, the next point addresses the issue.
- 2. Lexical Understanding of Non-Lexical Forms** Whether one believes a lexical form is a single concept or a grammatical construct formed from two or more independent smaller lexical concepts, the ultimate measure of a lexical knowledge base representation is whether it accounts for the meaning of the form. Thus, if one claims expressions such as “tax break” or “tax code” are not lexical concepts, then one is obligated to demonstrate how they can be understood through the heuristic combination of the lexical knowledge base representations for “tax”, “break” and “code”. It is also interesting that at this level we are not specifically addressing the combination of any specific type of language units. One could be talking about the English morphology and the understanding of terms such as “chocoholic” or “womanhood” just as readily as the understanding of open compounds. The mechanism and the lexical knowledge base elements available are the issue here.

## References

- [1] Amsler, R., “The Structure of the Merriam-Webster Pocket Dictionary”, Ph. D. Thesis, University of Texas at Austin, Austin, TX, 1980.
- [2] Amsler, R., “Computational Lexicology: A Research Program”, in *AFIPS Conference Proceedings: 1982 National Computer Conference*, American Federation of Information processing Societies, Arlington, VA, 1982, 657-663.
- [3] Amsler, R., “Words and Worlds”, in *Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing (TINLAP3)*, New Mexico State University at Las Cruces, NM, January 7-9, 1987.
- [4] Amsler, R., J. White, “Development of a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-Readable Dictionaries”, Linguistics Research Center, University of Texas at Austin, Final Report on NSF Project MCS77-01315, July 1979.
- [5] Botha, R., *Morphological Mechanisms: Lexicalist Analysis of Synthetic Compounding*, Pergamon Press, Oxford, England, 1984.
- [6] Bresnan, J., *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, MA, 1983.
- [7] Carroll, J., *What's in a Name*, W.H. Freeman, New York, 1985.

- [8] Cruse, D., *Lexical Semantics*, Cambridge University Press, Cambridge, England, 1986.
- [9] Flexner, S., *I Hear America Talking: An Illustrated History of American Words and Phrases*, Simon and Schuster, New York, 1976.
- [10] Isitt, D., *Crazic, Menty and Idiotal*, Acta Universitatis Gothoburgensis, Goeteborg, Sweden, 1983.
- [11] Leonard, R., *The Interpretation of English Noun Sequences on the Computer*, North-Holland, Amsterdam, 1984.
- [12] Levi, J., *The Syntax and Semantics of Complex Nominals*, Academic Press, New York, 1978.
- [13] Meys, W., *Compound Adjectives in English and the Ideal Speaker-Listener*, North-Holland, Amsterdam, 1975.
- [14] Michiels, A., “Exploiting a Large Dictionary Data Base”, Ph. D. Thesis, University of Liege, Liege, Belgium, 1981.
- [15] Peterson, J., “Webster’s Seventh New Collegiate Dictionary: A Computer-Readable File Format”, TR-196, Dept. of Computer Sciences, Univ. of Texas at Austin, May 1982.
- [16] Peterson, J., “Webster’s Seventh New Collegiate Dictionary: A Computer-Readable File Format”, Microelectronics and Computer Technology Corporation, Austin TX, 1987.
- [17] Quillian, R., “Semantic Memory”, in *Semantic Information Processing*, MIT Press, Cambridge, MA, 1968, 227-270.
- [18] Reichert, R., J. Olney, J. Paris, “Two Dictionary Transcripts and Programs for Processing Them. Volume I: The Encoding Scheme, PARSENT and CONIX”, TM-3978/001/00, System Development Corporation, Santa Monica, CA, 15 June, 1969.
- [19] Robins, G., “The NIKL Manual”, The Knowledge Representation Project, Information Sciences Institute, Marina Del Rey, CA, 1986.
- [20] Rusiecki, J., *Adjectives and Comparison in English: A Semantic Study*, Longman, London, 1985.
- [21] Shapiro, S., “The SNePS Semantic Network Processing System”, in *Associative Networks: Representation and Use of Knowledge by Computers*, Academic Press, New York, 1979, 179-203.
- [22] Sherman, D., “A New Computer Format for “Webster’s Seventh Collegiate Dictionary”, *Computers and the Humanities*, 8, 1974, 21-26.

- [23] Simmons, R., "Semantic Networks: Their Computation and Use for Understanding English Sentences", in *Computer Models of Thought and Language*, W.H. Freeman & Co., San Francisco, CA, 1973, 63-113.
- [24] Sowa, J., *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley, Reading, MA, 1984.
- [25] Walker, D., R. Amsler, "The Use of Machine-Readable Dictionaries in Sublanguage Analysis", in R. Grishman, R. Kittridge, (eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, 1986, Chapter 5, 69-83.
- [26] Warren, B., *Semantic Patterns of Noun-Noun Compounds*, Acta Universitatis Gothoburgensis, Goteborg, Sweden, 1978.
- [27] Wittenburg, K., "Natural Language Parsing with Combinatory Categorial Grammars in a Graph-Unification-Based Formalism", Ph. D. Thesis, University of Texas at Austin, Austin, TX, 1986.

# Discovering Relationships among Word Senses

Roy J. Byrd

IBM Thomas J. Watson Research Center

*e-mail: byrd@watson.ibm.com*

## Abstract

The direction that future development of natural language processing systems will take is toward increased semantic capability. Systems will need to behave as though they understand the texts that they process. An important prerequisite for this type of behavior is the creation of lexical knowledge bases in which word senses are clearly identified, endowed with appropriate lexical information, and correctly related to one another. This paper discusses issues that arise in creating such a knowledge base, paying particular attention to the process of discovering relationships among the word senses it contains.

## 1 Introduction

The Lexical Systems project at IBM Research has embarked on a long-term project to build CompLex, a computational lexicon which will be able to provide enhanced semantic information to natural language processing systems. A major goal of the project is to identify and extract information about word senses from machine-readable dictionaries. That is the subject of this paper. Equally important goals are to find such information by analyzing large text corpora and to develop automated sense disambiguation techniques based on dictionary and corpus analysis. These goals will not be discussed here.

Section 2 describes CompLex and its motivations. Section 3 provides detail about the identification and processing of word senses. Sections 4 and 5 describe relationships among word senses and give detailed descriptions of procedures for discovering some of those relationships.

## 2 CompLex: A Lexical Knowledge Base

CompLex is a network of word senses. It can be viewed from two different perspectives: as a computational lexicon for use by NLP systems and as a prototype lexical knowledge base leading ultimately toward more general knowledge processing technology for artificial intelligence.

The original motivation for building CompLex came from the first perspective. NLP systems suffer from having lexicons which handle polysemous words inadequately. Some systems (Grishman and Kittredge 1986, Nirenberg and Raskin 1987) approach the problem of polysemy by restricting their application domain to a subworld in which

most words have only one meaning. In these systems, ambiguous words are assigned a single meaning – usually the most frequent sense for the domain – and only the lexical information for that sense is stored. For example, the lexical entry for the word *file* in a business application would be treated as a repository for information, and never as a tool for smoothing surfaces. Other systems (Heidorn, et al. 1982, Byrd 1983) deal with polysemous words by combining information from multiple senses in various ways. For example, IBM's UDICT dictionary (Byrd 1983, Klavans and Wacholder 1987) marks certain features on a word only if all of its senses bear that feature. Thus, although *prince* is marked [+human], *king* is not, because [+human] is inappropriate for the sense of *king* which refers to the chess piece.

The decision to create a computational lexicon that handles polysemy properly leads immediately to several other decisions. First, criteria for separating word senses must be selected and used to find individual senses. Next, disambiguation strategies for selecting the appropriate sense of a polysemous word used in context must be devised. Information necessary to support the disambiguation logic must be found and stored in the lexicon. In addition to homography, which is a rather specialized kind of inter-sense relationship along which disambiguation operates, there are more general relationships among word senses that must also be stored. These include grammatical relationships such as selectional restrictions as well as more semantic ones like synonymy or hypernymy.

The picture that emerges from these considerations is one of CompLex as a large network of nodes representing word senses and containing information that is inherent to those senses. The links among the senses appear as arcs having labels such as "homograph", "synonym", "hypernym", "typical object", etc. This model has much in common with the semantic networks that are commonly used to represent knowledge in artificial intelligence applications. In fact, a CompLex-like network can offer a real benefit to artificial intelligence: it is a naturally occurring example of a large coherent network to which the techniques developed for representing, storing, and processing knowledge may be applied. For this reason, we can also view CompLex as a lexical knowledge base. Furthermore, if we can extrapolate from word senses to concepts, then it may be possible to use experience gained in building and processing CompLex (for example the development of inferencing mechanisms and rules) to advantage when dealing with more general knowledge bases.

### 3 Word Senses

The advantages claimed for CompLex in the previous section can only be realized if we can effectively identify large numbers of word senses and assign inherent information and relationships to them. In this section, we describe word senses and their information in general terms. The next two sections give details about specific inter-sense relationships and discuss techniques for finding them.

Work in computational lexicology has recently concentrated on extracting lexical information from machine-readable dictionaries and text corpora. Byrd, et al. (1987) and references given there provide the background for this method of working. Our CompLex research is based on these same tools and methodologies. In particular, our search for word senses begins with the senses listed in the machine-readable dictionaries

that we process. We recognize that no single dictionary's sense distinctions are entirely "correct" or appropriate for the uses to which CompLex will be put. In fact, the necessity to deal with different views about how semantic space is carved into word senses gives rise to a "mapping problem" which will be described shortly. Nevertheless, being able to begin our work with word senses already identified gives us tremendous leverage.

We are convinced that it will be easier to extract and, where necessary, adjust the senses already identified by lexicographers than it would be to invent senses from scratch or to reproduce the work that lexicographers have already done. Furthermore, in most dictionaries, word senses come endowed with the kinds of information that we ultimately want to have in CompLex. Thesaurus type dictionaries already have lists of synonyms. Definitions in monolingual dictionaries give us hypernyms as well as information about arguments that words take. Bilingual and learners' dictionaries are frequently very explicit in the kinds of grammatical information that they present. It is crucial that this information is attached not just to words, but to the word senses recorded in these dictionaries. Even though the senses we begin with may not be ideal, if we can map the published senses onto the ultimate CompLex senses, then the wealth of lexical information available can be assigned to those senses.

We expect to derive the CompLex word senses which will contain all this information by a process of comparing the senses found in different machine-readable dictionaries and striking a compromise among them. This leads to the mapping problem, mentioned earlier, which involves determining the "best" mapping from the senses given for a word in one dictionary to the senses given for the same word in another. For a variety of reasons, the sense distinctions found in published dictionaries are less than ideal. Monolingual dictionaries for native speakers tend to give too many senses involving subtle distinctions and rare or obsolete uses. Bilingual dictionaries often do not disambiguate a polysemous source language word if the target word is ambiguous in the same way. Synonym dictionaries tend to have few entries for concrete nouns.

Despite these difficulties, we believe it will be possible to map senses from one dictionary using automated and semi-automated mechanisms. The process of achieving these mappings and the fact of having them for inspection should give us insights into the nature of the ideal senses that we will subsequently define for CompLex. We take very seriously the notion (advocated by Atkins and Levin 1988) that only good linguistic analysis will enable the definition of ideal senses to which published dictionary senses can be successfully mapped. The techniques that we develop while attempting bilateral dictionary mappings will be important ingredients of that linguistic analysis, and the analysis will help us develop and refine the techniques.

The mapping procedure that we are building is based on the construction of "dictionary sense property vectors". These vectors are lists of labeled items where each item is a word sense property that is derivable from the information given for that sense by the dictionary. The following is a partial list of the types of information that we expect to have in the vectors:

- part-of-speech
- the words in a definition, perhaps morphologically analyzed into lemmata or derivational bases
- the dictionary sense number

- the hypernym (genus term)
- subject selectional restrictions
- object selectional restrictions
- typical subject
- typical object
- typical instrument
- manner
- purpose
- material
- synonyms

Values for many of these properties can be obtained simply by querying the lexical data bases for the dictionaries involved. These lexical data bases have been described in Neff and Boguraev (1989). Some of the less explicitly represented properties will require more subtle combinations of queries and syntactic analysis of definitions and examples for their discovery. Figure 1 shows that many of these properties are available in the definitions from *Webster's Seventh New Collegiate Dictionary*. More detailed descriptions of the extraction of properties like these will be presented later in the paper.

It is clear that no single sense property vector will have values for all of these properties. Monolingual dictionaries often do not give synonyms, and bilingual dictionaries usually do not give definitions, for example. Nevertheless, there will be enough information for the kind of mapping procedure that we envision. The procedure for mapping a word begins by computing a set of sense property vectors for entries from each of the two dictionaries being mapped. Next, a set of heuristics computes an "optimal" mapping of the two sets onto one another. A simple case in which this procedure yields extremely good results involves mapping two synonym dictionaries. When *The New Collins Thesaurus* and *Roget's II: The New Thesaurus* were mapped, the sense property vectors consisted only of the synonyms listed for each sense (plus sense number and part-of-speech). In this case, the mapping heuristic was to compute the overlaps among the synonym lists and to choose the mapping in which the overlaps were maximized. This strategy uses the same synonym intersection logic that has been used for sense disambiguation of the cross references in the *Collins Thesaurus* and was reported in Chodorow, Ravin, and Sachar (1988). A similar disambiguation procedure has been described in Lesk (1986). Intersection was also used to partially map the two sides of the *Collins-Robert English-French* and *French-English* dictionaries, as part of a disambiguation study reported in Rizk (1989).

A similar word intersection scheme has been used in an attempt to map between dictionaries containing definitions. In this case, intersections were taken of the morphologically normalized words in definitions from *Webster's Seventh*, *Longman Dictionary of Contemporary English (LDOCE)*, and *Roget's II* (which contains definitions as well

genus
<b>reel:</b> <i>n</i> a revolvable device on which something flexible is wound
typical subject
<b>acrocarpous:</b> <i>a</i> (of a moss) having the archegonia and hence the capsules terminal on the stem
typical object
<b>mangle:</b> <i>vt</i> to press or smooth (as <i>damp linen</i> ) with a mangle
instrument
<b>mangle:</b> <i>vt</i> to press or smooth (as damp linen) with a <i>mangle</i>
manner
<b>wallop:</b> <i>vt</i> to hit with force
object selection
<b>single:</b> <i>vt</i> to select or distinguish (a <i>person</i> or <i>thing</i> ) from a number or group
subject selection
<b>blue:</b> <i>a</i> (of a woman) LEARNED, INTELLECTUAL
purpose
<b>file</b> <i>n</i> a tool usu. of hardened steel with cutting ridges for <i>forming</i> or <i>smoothing</i> surfaces.

Figure 1. Elements of dictionary sense property vectors, from Webster's Seventh definitions.

as synonym lists). This procedure showed limited success, with fewer than 50% of the resulting mappings being correct. The failures are primarily of two types; incorrect mappings were proposed in some cases and correct mappings were missed in others. These failures reveal that the lists of undifferentiated definition words are not selective enough for adequate mapping. However, inspection of these results suggests that our overall plan for mapping shows promise. When the words in the definitions are assigned to specific sense properties (as are the italicized words in Figure 1), and the matching is done property-by-property, the mappings will be based on more constrained criteria. This work is at its beginning and will be reported on later.

Once successful bilateral mappings can be achieved, we will be in a position to define the criteria for sense distinctions that will motivate the "ideal" senses that we store in CompLex. Experience gained by working with early versions of these ideal senses will allow us to refine our mapping procedures, in an iterative fashion. First of all, we will then know which sense properties are most effective in identifying senses. We will also have a better idea of what role the sense properties we are able to identify (for the vectors) will play in the applications that CompLex will serve. It should be clear that these properties are precisely the items of lexical information that we want to store in CompLex. Once criteria for defining CompLex senses have been established, we can map from individual machine-readable dictionaries onto the CompLex senses and transfer the lexical information directly.

## 4 Relationships among Word Senses

The information recorded in dictionary sense property vectors consists mainly of information about relationships from word senses to other words. The sources of these relationships are word senses, namely those to which the property vectors belong. The

targets to which the relationships point, however, are non-disambiguated words that appear in the source words' entries. If these targets can be further refined to word senses, then these relationships are precisely what we want to store in CompLex. Among the relationships that are of initial interest in our work are the following.

- homography
- hypernymy (genus) and hyponymy
- synonymy
- typical arguments
- selectional restrictions
- miscellaneous (e.g., *goods sold by merchants*)

The next section of the paper describes procedures for finding the target references of these relationships and for refining the word references into sense references, where possible.

## 5 Finding Relationships among Word Senses

In principle, dictionary sense property vectors can be viewed as being open-ended; the kinds of lexical relations that can be extracted from dictionary entries for use in mapping their senses is limited only by our ability to analyze those entries with the tools we have. In fact, workers on dictionary based semantic disambiguation (Jensen and Binot 1987 and Yael Ravin, personal communication) stress the notion that there is no a priori limit on the types of usable information that can be obtained from machine-readable dictionaries.

We have also pointed out that certain of the properties are also exactly the ones that CompLex will need to store. It might be imagined that we would seek to capture only those properties that are fairly regular and densely available from machine-readable dictionaries. However, it is important to note two things. First, the analysis techniques that we employ are never 100% effective. Our techniques will always miss something, either because a certain type of information was not exhaustively recorded by the lexicographers or because its representation is not entirely consistent. Second, as we gain experience with the use of CompLex in specific applications, we may be willing to accept more "interesting" but less complete information. After listing our tools, we present a range of lexical relationships that we have attempted to extract.

### 5.1 Tools

The tools used in this research have been adequately described elsewhere (Byrd, et al. 1987, Neff, et al. 1988, Neff and Boguraev 1989), and will only be listed here. Our machine-readable dictionaries are stored in lexical data bases and accessed with LQL, a lexical query language. A great deal of inherent information about words is stored in the UDICT computational lexicon which also performs morphological analysis. The

TUPLES text analysis system finds, counts, lists, and concordances individual words and phrases in large amounts of text. We use the PLNLP English Grammar (PEG), described in Jensen (1986) to parse dictionary definitions and examples. A syntactic pattern matching tool, APT, is used to locate words in PEG parses that fill specific syntactic roles.

## 5.2 Homography

When two word senses are named by the same word, they are “homographs”. This usage of the term is somewhat different from the customary one in lexicography, where homographs are really different words that happen to be spelled the same and where the senses of an ambiguous word are not called homographs. For computational treatment, this distinction is subordinated to the fact that the same character string is used to access all of the senses that we are calling homographs.

From one point of view, this relationship is trivial to capture. We merely must find all the senses that dictionaries associate with a given word and link those senses with homograph links. The difficulty comes when we try to exploit these links. Their exploitation occurs during the sense disambiguation process. Analysis of that process reveals that the homograph relationship does not consist of simple links between word senses. Rather the link must include information about how to choose among the competing word senses when the word is seen in context. Thus, the homograph links are the homes of the semantic-ID tags described in Atkins (1987). As Atkins points out, semantic-ID tags can only be discovered by analyzing text corpora. Since we restrict this discussion to information discoverable from dictionaries, semantic-ID tags will not be further discussed here.

## 5.3 Hypernymy/Hyponymy

Hypernymy is known to workers in artificial intelligence as the “ISA” relationship. Lexicographers usually call it the “genus” relationship. Usually, a word’s hypernym is the name of the concept that is superordinate to the concept named by the word we started with. The collection of all hypernym links available from a dictionary produces a fairly complete taxonomy of concepts which can be used for many purposes. Such taxonomies, which can only be built for nouns and verbs, have been described in Amsler (1980) and Chodorow, et al. (1985).

A problem with previous work on dictionary-based taxonomies has been that the nodes in them are words, not word senses. We are currently building a sense-disambiguated taxonomy in which the nodes are word senses. The work is being done with respect to senses given by *Webster’s Seventh*. However, we have already argued that if dictionary senses can be mapped onto CompLex senses, then information from the dictionary can also be mapped. In the case of hypernymy, once the superordinate (hypernym) and the subordinate (hyponym) of every instance of the relationship is mapped, we will have a complete taxonomy with respect to CompLex senses.

The procedure for disambiguating the hyponyms is fairly straightforward. As in the previously reported work, we locate the syntactic heads of the noun phrase and verb phrase definitions. In this work, we use the PEG parser and the APT pattern

matcher to isolate the syntactic head, rather than the heuristics used in the 1985 version. Consequently, for each hypernym-hyponym pair, the sense number of the hyponym is automatically available (since it was stored with the definition which was parsed) and the hyponyms are disambiguated.

Disambiguation of the hypernyms will require a collection of techniques. Hypernyms that are monosemous (having only one sense) can be marked instantly. However, although monosemous words account for two-thirds of the words in the dictionary, they tend not to be used as hypernyms. A second approach is to use our mapping logic to choose the sense(s) of a polysemous hypernym that best map(s) to each disambiguated hyponym. We have tested this notion with the current intersection-based mapping algorithm and have obtained mixed results. We expect that improvements planned for the mapping program will also improve our hypernym disambiguation performance.

## 5.4 Synonymy

We have developed a set of techniques for disambiguating the cross references in standard synonym dictionaries. Chodorow, et al. (1988) point out that the assignment of sense numbers to cross references can be performed (1) by looking for symmetric cross references and (2) by intersecting the synonym lists associated with each sense and choosing the senses that have the largest intersections. Procedures based on these techniques have been used to disambiguate over 80% of the cross references in the *Collins Thesaurus*.

*Roget's II* is a synonym dictionary which has definitions associated with its senses (i.e., each sense consists of a definition and a synonym list). Since the lexicographers were careful to re-use the same definitions to mark the corresponding senses of synonymous words, the dictionary is in effect already sense disambiguated. It was a straightforward matter to process the machine-readable *Roget's II* in order to mark sense numbers on all the cross references (Michael Gunther, personal communication).

The synonym list intersection method has been used to create a procedure which maps *Collins Thesaurus* and *Roget's II* onto one another. Again, this procedure is straightforward and quite effective. The interesting thing is that the resulting (augmented) synonym lists are paired with the definitions from *Roget's II*. This now opens the way to using our sense mapping logic to map these large sets of synonyms onto monolingual English dictionaries, and hence onto CompLex. In fact, since all of our monolingual and bilingual dictionaries contain at least some synonyms, the chances of mapping the combined thesauruses with a monolingual dictionary are fairly high.

## 5.5 Typical Arguments and Selectional Restrictions

Many words impose restrictions on the kinds of arguments they take when used in context. The argument positions usually so constrained are the subjects and objects of verbs and the modified nouns of adjectives. In addition, positions like instrument, goal, source, location, etc. may also restrict the kinds of fillers they allow. Dictionaries often explicitly list these constraints in their entries. Figure 1 shows that typical arguments and selectional restrictions are given in *Webster's Seventh* as the values of parenthetical expressions. It is easy to see how typical objects of verbs, for example, can be extracted

from such definitions. We must merely take the value of the parenthetical expression if (1) it is a noun phrase and (2) it is the object of the head of the verb phrase definition. Other similar conventions are used for typical subject, etc. The following is a small example of the results of applying this procedure to verbs starting with "a" in *Webster's Seventh*.

accredit	an educational institution
accredit	an envoy
adopt	a child of other parents
amputate	a limb or projecting part
anodize	a metal
appeal	a case
arrange	a musical composition

From such output, it is easy to extract the typical object relationship between the sense of the head verb and the noun which is the head of the object noun phrase. In order to convert the noun reference into a particular sense reference, we must solve a problem analogous to the problem of disambiguating hypernyms. No work has been done on this problem beyond what has already been mentioned.

It is interesting to consider how these typical argument relationships might be used by applications that incorporate CompLex. In the sense of *arrange* that takes *a musical composition* as its object, for example, an application program might be prepared to accept words like *symphony*, *song*, *capriccio*, *sonata*, etc. In this case, the restriction is not to a single word (sense) but to a set of hyponyms of (*musical*) *composition* in the noun taxonomy. Of course, the relationship is not always strict hyponymy; for *appeal*, suitable objects might be *verdict*, *punishment*, *decision*, etc., in addition to *case*. Nevertheless, it appears that appropriate arguments for argument-taking words can be obtained using relationships (like hyponymy, synonymy, etc.) that we plan to keep in CompLex.

This point of view suggests that selectional restrictions are the same thing as typical argument designations, the only difference being that for selectional restrictions the argument can only be words like *thing*, *animal*, or *person*, i.e., words that are near the top of the noun taxonomy. Thus the following equivalences can be used:

human obj	= typical object: person, someone, another
animate obj	= typical object: animal
concrete obj	= typical object: thing, something

Applying this convention to verbs from *Roget's II*, we can obtain a list of several hundred verbs taking human objects, including *abash*, *accompany*, *alert*, *bring around*, *bump off*, *chagrin*, and *chase*. The convention applies equally well to *Webster's Seventh* and *LDOCE*. In addition, selectional restrictions (as well as other types of typical arguments) can be obtained from the bilingual dictionaries, where they are represented in a slightly different format. Figure 2 shows several examples from among the almost 600 verbs marked as taking human objects in the *Collins-Robert English-French Dictionary*.

**assuage** *vt [person]* apaiser, calmer.  
**castigate** *vt [person]* châtier, corriger, punir.  
**herd together** *vt [people]* rassembler en troupeau.  
**laugh at** *vt [person]* rire de, se moquer de.  
**pacify** *vt [person]* calmer, apaiser.  
**station** *vt [people]* placer, mettre, poser.

Figure 2. Verbs marked for *human-object* in the Collins-Robert English-French Dictionary.

**antiquarian:** a person  
    who studies, collects, or sells  
        objects that are very old

**apothecary:** one  
    who prepares and sells  
        drugs or compounds for medicinal purposes

**blockbuster:** a person  
    who gets white people to sell  
        their houses to him cheaply by telling them that black  
            people are going to move into the area

**confectioner:** a person  
    who makes or sells  
        sweets, ice cream, cakes, etc.

**estate\_agent:** a person whose business is to bring together people  
    who want to sell  
        and people who want to buy houses, property, or land, and to  
            look after the property of others

**florist:** one  
    who sells  
        flowers and ornamental plants

**newsboy:** a boy or man  
    who sells  
        or delivers newspapers

**pardoner:** (800 to 500 years ago) a person  
    who went about the country selling  
        official religious INDULGENCES

**pork\_butcher:** a dealer in meat  
    who sells  
        (only) PORK or products made from it, such as SAUSAGES

**pusher:** A person  
    who sells  
        narcotics illegally.

Figure 3. Concordance of "who sell", showing goods sold by merchants.

## 5.6 Miscellaneous Relationships: “goods sold by merchants”

As noted at the beginning of this section, it may sometimes be desirable to obtain specialized lexical relations that are not exhaustively represented in any of the machine-readable dictionaries. In such cases, it is useful to be able to combine results from as many sources as possible. We have already seen that various forms of the typical argument relations can be obtained from all of the monolingual and bilingual dictionaries. Figure 3 shows an intermediate stage in the derivation of a more specialized relation, namely the one that links merchants to the goods that they typically sell.

This derivation began with the extraction (from *Webster’s Seventh*, *LDOCE*, and *Roget’s II*) of nouns whose definitions contain a form of the word *sell*. The resulting definitions were subjected to a number of analyses, one of which yielded the concordance shown in Figure 3. This concordance resulted from a TUPLES request for words whose definitions contain the collocation “who sell” within a window of 7 words. The goods sold are marked with italics in the figure. There are several things to notice about this data. First, the names of the merchants are given as the head nouns being defined. Next, the objects typically sold appear as the direct object of the verb *sell*. However, the task of finding goods sold by merchants is not quite that simple. The definitions for *blockbuster* and *estate agent* (neither of whom sells houses) demonstrate that the relative pronoun *who* must introduce a relative clause that modifies the genus term of the definition. The definition of *confectioner* shows that the verb *sell* and the object nouns may both be conjoined. Our analysis procedures will need to be able to sort out such syntactic structures.

This example clearly demonstrates the need for the full range of tools that we use in our research. The combination of lexical data bases and query language, text analyzer, parser, and syntactic structural analysis makes it possible to develop quite abstract and interesting lexical properties. Without such tools, our research goals would have to be much more modest.

## 6 Conclusions

This paper has described aspects of an ongoing research project devoted to constructing a large lexical knowledge base. The focus has been on techniques for finding relationships among word senses in that knowledge base. An important observation is that the same lexical information that serves to identify word senses, using the property vector mechanism, is also the information that we want to store in CompLex. This fact increases the utility of our discovery procedures by letting us use their results in multiple ways.

The lexical knowledge base that we envision is a modest goal by comparison with the world knowledge bases sought by projects such as EDR (Electronic Dictionary Research) in Japan or CYC at MCC. However, unlike those projects which rely on a large amount of manually input information, the CompLex project attempts to derive most of its information by automated analysis of dictionaries and texts. We believe that a good lexical knowledge base is a subset of, rather than disjoint with, a world knowledge base. If we succeed, therefore, CompLex may provide a good alternative starting point from which to construct a world knowledge base.

## Acknowledgments

In this paper, the pronoun “we” is often used in references to work performed and beliefs held as part of this research. This is not the imperial “we”! It refers quite concretely to the members of the Lexical Systems project at IBM Research. They are Bran Boguraev, Martin Chodorow, John Justeson, Slava Katz, Judith Klavans, Mary Neff, Yael Ravin, Omneya Rizk, Zofia Roberts, and Nina Wacholder. They know their individual contributions. I am grateful to them all. I also thank Sue Atkins for many discussions about the nature of lexicography and the pitfalls of relying on published dictionaries. If the research described in this paper should fail, it will not be her fault.

## References

- [1] Amsler, R. A., “The Structure of the Merriam-Webster Pocket Dictionary,” Doctoral Dissertation, TR-164, University of Texas, Austin, Texas, 1980.
- [2] Atkins, B. T., “Semantic ID tags: Corpus Evidence for Dictionary Senses,” in *The Uses of Large Text Databases: Proceedings of the Third Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, 1987, 17-36.
- [3] Atkins, B. T., and B. Levin, “Admitting Impediments,” in *Information in Text: Proceedings of the Fourth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, 1988, 97-113.
- [4] Byrd, R. J., “Word Formation in Natural Language Processing Systems,” in *Proceedings of IJCAI-VIII*, 1983, 704-706.
- [5] Byrd, R. J., N. Calzolari, M. S. Chodorow, J. L. Klavans, M. S. Neff, and O. A. Rizk, “Tools and Methods for Computational Lexicology,” *Computational Linguistics*, 13 (3-4), 1987, 219-240.
- [6] Chodorow, M. S., R. J. Byrd, and G. E. Heidorn, “Extracting Semantic Hierarchies from a Large On-line Dictionary,” in *Proceedings of the Association for Computational Linguistics*, 1985, 299-304.
- [7] Chodorow, M. S., Y. Ravin and H. E. Sachar, “A Tool for Investigating the Synonymy Relation in a Sense Disambiguated Thesaurus,” in *Proceedings of the Second ACL Conference on Applied Natural Language Processing*, Austin, Texas, 1988, 144-151.
- [8] Grishman, R. and R. Kittredge, *Analysing Language in Restricted Domains*, Erlbaum, 1986.
- [9] Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow, “The EPISTLE Text-Critiquing System,” *IBM Systems Journal*, 21 (3), 1982, 305-326.

- [10] Jensen, K., "Parsing strategies in a broad-coverage grammar of English," IBM Research Report RC12147, IBM T. J. Watson Research Center, Yorktown Heights, New York, 1986
- [11] Jensen, K. and J.-L. Binot, "Disambiguating Prepositional Phrase Attachments by Using On-line Dictionary Definitions," *Computational Linguistics*, 13 (3-4), 1987, 251-260.
- [12] Klavans, J. L. and N. Wacholder, "Documentation of Features and Attributes in UDICT," IBM Research Report RC14251, IBM T. J. Watson Research Center, Yorktown Heights, New York, 1987.
- [13] Lesk, M., "Automatic Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," in *Proceedings of the 1986 SIGDOC Conference*, 1986.
- [14] Neff, M. S., R. J. Byrd, and O. A. Rizk, "Creating and Querying Hierarchical Lexical Data Bases," in *Proceedings of the Second ACL Conference on Applied Natural Language Processing*, Austin, Texas, 1988, 84-92.
- [15] Neff, M. S., and B. K. Boguraev, "Dictionaries, dictionary grammars, and dictionary entry parsing," in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 1989, 91-101.
- [16] Nirenberg, S. and V. Raskin, "The subworld concept lexicon and the lexicon management system," *Computational Linguistics*, 13 (3-4), 1987, 276-289.
- [17] Rizk, O. A., "Sense Disambiguation of Word Translations in Bilingual Dictionaries: Trying to Solve the Mapping Problem Automatically," master's thesis, Courant Institute of Mathematical Sciences, New York University, New York, New York, 1989; also available as IBM Research Report RC14666, IBM T. J. Watson Research Center, Yorktown Heights, New York, 1989.

## Dictionaries

- [18] *Collins-Robert English-French Dictionary*, Collins, London and Glasgow, 1978.
- [19] *Longman Dictionary of Contemporary English*, Longman, Harlow, England, 1978
- [20] *The New Collins Thesaurus*, Collins, London and Glasgow, 1984.
- [21] *Roget's II: The New Thesaurus*, Houghton Mifflin, Boston, 1980.
- [22] *Webster's Seventh New Collegiate Dictionary*, G. & C. Merriam, Springfield, Massachusetts, 1963.

# Machine Readable Dictionary as a Source of Grammatical Information

Eva Hajčová

Institute of Formal and Applied Linguistics

Alexandr Rosen

Institute of Theoretical and Computational Linguistics

Charles University, Prague

e-mail: *hajicova@cspguk11.bitnet*

## Abstract

The present contribution describes an enterprise in collecting lexical data for an English parser in the context of a bilingual research project. The primary source of grammatical information is a computer usable version of OALD (Hornby, 1974). The target lexicon's structure of verbal valency frames, inspired by the theoretical framework of functional generative description, includes an underlying level. Its content can be derived under some human supervision from OALD's verb pattern codes. Results confirm the usefulness of machine readable dictionaries for NLP applications.

## 1 Introduction

Though some of the older Machine Translation (MT) projects concentrated chiefly on a build-up of effective grammatical modules of the systems, it was soon clear that the issues connected with adding new lexical items into the systems were far from trivial. The inquiries into large lexical databases and machine readable text corpora represented an invaluable support not only for a particular domain of natural language processing (NLP) but for linguistic research as such. Don Walker belongs to the first who realized this; his research and organizational efforts contribute a great deal to the considerable progress in this domain we are witnessing (see, e.g., Walker and Amsler, 1986; Walker, 1987; Walker et al., 1987; Walker, 1989).

In the present contribution, we first delineate the position of our current project in machine translation with regard to the discussion between the so-called empiricists and rationalists (Section 2), passing over to a brief characterization of grammatical data we assume the lexicon should include (Section 3). Finally (in Section 4) we describe a process of constructing a reasonably large lexicon (monolingual, for the time being) from some freely available machine-readable dictionaries. As an attempt to make the most of grammatical information present in one of the sources we suggest a procedure for the derivation of underlying valency frames from verb pattern (complementation) codes in the source (an electronic version of the *Oxford Advanced Learner's Dictionary*

of *Current English* – OALD, Hornby, 1974). The intended use of the lexicon is within a parsing/tagging and bilingual comparison tool, which will provide empirical data for the transfer module of a machine translation system, but a variety of other purposes can be imagined.

## 2 Empirical vs. Rationalist Paradigms

In recent years, most prominent linguistic research groups have turned towards large-scale resources of linguistic data, such as computerized lexicons and machine-readable collections of text corpora, to have a more immediate access to empirical data supporting their theoretical research, and to speed up the development of NLP applications. This trend has led to a methodological dispute in the field of machine translation between the so-called empiricists and rationalists (see TMI-92), or, in other words, between statistical and analytical approaches. The project of MT between English and Czech called MATRACE and designed in the years 1990-91 by the research team of computational linguistics at Charles University, Prague, represents an intermediate approach (Hajič et al., 1992): it applies a classical transfer paradigm, with “rationalist” modules of analysis and synthesis, and with “empiricist-based” transfer building on correspondences established at an underlying level (for some other intermediate approaches, see, e.g., Grishman, Kosaka, 1992; Lehmann, Ott, 1992).

## 3 The Need of a Rich Lexicon

The shape of this underlying level builds upon the theoretical framework of the so-called functional generative description (for its most comprehensive account, see Sgall, Hajičová, Panevová 1986), one of the distinguishing features of which is the representation of the syntactic structure in terms of dependency trees. The dependency approach to syntax is characterized *inter alia* by the fact that a substantial amount of grammatical information is included in the lexicon: this concerns first of all the assignment of “frames” (called case frames, theta grids, valency slots in different theories) to individual lexical entries. An attempt to extract valency frames for verbal entries from subcategorization codes in the computer-readable version of OALD is described in the next section; for a discussion and substantiation of the repertoire of valency relations and their classification we refer to Chapter 2 of the above mentioned book by Sgall et al. (1986).

But of course, there are never enough lexical data: apart from the information on the valency of verbs, valency of other word classes as well as information on co-occurring lexical items should be present. Furthermore, in an MT project, the factor of primary importance is the quality and volume of contrastive lexical data. To accumulate reasonably large monolingual and bilingual resources, various approaches are pursued, including automatic extraction of lexical data from ordinary, tagged and parallel texts.

## 4 From Lexical Resources to a Structured Lexicon

### 4.1 Converting and Integrating the Resources

Our primary source of English lexical data is the CUVOALD, or the “Expanded Computer Usable Version” of the *Oxford Advanced Learner’s Dictionary of Current English*, 3rd edition (OALD, Hornby, 1974), which is available from the Oxford Text Archive (see Mitton, 1986). CUVOALD lists all headwords, headword variants and derivatives with simple codes denoting word classes and inflection patterns, supplemented by verb pattern codes for verbs. See Figure 1 for some sample entries.

spell	spel	J0, K6	2A, 6A, 15B
spellbinder	'spelbaInd@R	K6	
spellbound	'spelbaUnd	OA	
speller	'spel@R	K6	
spelling	'spelIN	M6	
spelt	spelt	Hc, Hd, Jc, Jd, L@	2A, 6A, 15B
spend	spend	J5	2A, 6A, 14
spender	'spend@R	K6	
spending	'spendIN	Jb	2A, 6A, 14
spends	spendz	Ja	2A, 6A, 14
spendthrift	'spendTrift	K6	
spent	spent	Jc, Jd, OA	2A, 6A, 14

Figure 1: Sample entries from the Expanded Computer Usable Version of the Oxford Advanced Learner’s Dictionary of Current English

J0 stands for a regularly inflected transitive and intransitive verb (note, however, that there is an irregular variant ‘spelt’ for past tense and past participle), K6 for regularly inflected countable noun, OA for uninflected adjective, etc. The verb pattern codes (here 2A, 6A, 14 and 15B) correspond to the system used in Hornby (1974). Irregular words like ‘spend’ have all their forms listed. There are approximately 38,000 such entries, but homographs like ‘spell’ are frequently treated as a single entry.

First, this dictionary was processed to provide more explicit information about irregular forms and some verb patterns. Subsequently, it was merged with another source: a file called SOED or “Shorter Oxford English Dictionary” (Dolby et al., 1963), itself compiled from two sources, the Shorter Oxford proper and a Webster’s dictionary. The result has about 110,000 entries.

Except for word class codes the SOED file does not give much morphological or syntactic information, but it gives information about usage. The combination of usage codes from the two dictionaries integrated in the SOED file with the information on the presence or absence of a given word in all of the source dictionaries were employed to give a quasi-estimate of the word’s frequency. Every entry includes a Status item, whose value ranges from 1 to 17, depending on the above criteria. If Status=1, the entry was present in all of the sources, the usage codes in the SOED file being S for standard. If Status=7, the entry was present in OALD and absent in Webster, while SOED gave a non-standard usage code (i.e., archaic, obsolete, dialect, rare, etc.). If an entry was present only in SOED with a non-standard usage code, the result is Status=16. The Status information could be employed for a preferential treatment of ambiguities.

The result was expressed in a format used by the lexical component of ELU, a software environment for natural language processing developed at ISSCO, Geneva (see Russell et al., 1992). The format is transparent and flexible enough to allow portability of the target lexicon to other systems and formalisms. The ELU lexical component also supports multiple inheritance (entries inherit information from higher classes) and defaults (which can be overridden at lower levels), but these features are not fully employed in the lexicon to simplify potential conversions to other formats. Figure 2 shows the CUVOALD entries from Figure 1 after some modification and conversion into the ELU format, enriched by entries from the SOED file.

```

#Word spell          (DualPtPp CStemV TransInTrans Verb)
Pt=spelt Pp=spelt VP=[2A,6A,15B]
Status=1 Usage=[s,s]
(CStemN Count Noun)
Status=1 Usage=[s,s]
(CStemV Trans Verb)
Pt=spell-bound Pp=spell-bound
Status=10 Usage=[s,s]
(Adj) Status=14 Usage=[x,s]
(CStemN Count Noun)
Status=9 Usage=[x,x]
(NoInflJ Adj) Status=9 Usage=[x,x]
(CStemN Count Noun)
Status=1 Usage=[s,s]
(CStemN CountUncount Noun)
Status=5 Usage=[x,s]
(NoPlN Uncount Noun)
Status=1 Usage=[s,s]
(Noun) Status=10 Usage=[s,s]
(Verb) Status=14 Usage=[x,s]
(Adj) Status=10 Usage=[s,s]
(Noun) Status=15 Usage=[d,d]
(Noun) Status=10 Usage=[s,s]
(CStemV TransIntrans Verb)
Pt=spent Pp=spent VP=[2A,6A,14,19B]
Status=1 Usage=[s,s]
(Noun) Status=10 Usage=[s,s]
(Noun) Status=12 Usage=[r,s]
(CStemN Count Noun)
Status=1 Usage=[s,s]
(Noun) Status=14 Usage=[x,s]
(Noun) Status=10 Usage=[s,s]
(CStemN Count Noun)
Status=1 Usage=[s,s]
(Adj) Status=10 Usage=[s,s]
(NoInflJ Adj) Status=5 Usage=[x,s]

```

Figure 2: Sample entries from the compiled lexicon

The codes in parentheses refer to classes from which the lexical item inherits information (i.e., constraints given in feature structures). The order of the codes in the

parentheses is meaningful: a constraint in a previously referred class may override a constraint in a following class. The entries which were not present in CUVOALD (Status > 9) lack more detailed grammatical information. However, defaults seem to work in most of such cases. Some information retained from CUVOALD may not even be necessary, e.g., the information about regular inflection paradigms (CStemN = a noun whose stem ends with a non-sibilant consonant).

## 4.2 The OALD Verb Patterns

Whereas the derivation of lexical information from CUVOALD and SOED word class and usage codes was relatively straightforward, the OALD verb pattern codes present a real challenge. The dictionary classifies verbs according to the number and form of complements into 51 “verb patterns”, marked by numbers 1-25, supplemented in some cases by letters (4A, 4B, 4C, 4D, 4F). A pattern groups together verbs which exhibit the same behaviour in a standard context and are subject to the same set of transformations under specified conditions. A single pattern is also used for verbs which allow the same morphosyntactic variations of a complement. A different verb pattern is, however, used if only a subset of the relevant class permits the variation.

The classification of verb patterns provided by OALD is not without shortcomings; for our purpose, the most problematic seems to be a too surface-level definition of some verb patterns. These classes are then quite a heterogeneous collection: VP14 denotes verbs in all of the following uses, the only requirement being that the verb is followed by a noun and a prepositional phrase:

*They accused him of stealing the book.  
I explained my difficulty to him.  
Compare the copy with the original.*

Another “misbehaved” pattern is VP4A where, depending on the verb, the infinitive can be a complement or an adjunct:

*The swimmer failed to reach the shore.  
He came to see that he was mistaken.  
She stood up to see better.*

## 4.3 Valency Information in the Target Lexicon

The target lexicon (i.e., the lexicon used for the MT project) contains the following information about the valency of a verb (or its complementation), grouped in an entry as a complementation paradigm:

SUBCATEGORIZATION LIST (SC) gives syntactic and morphological categories for every dependent, i.e. either a “participant” (complement, which may be obligatory or optional) or an “obligatory free modification” (obligatory adjunct). An item in the list is in fact an underspecified representation of the corresponding dependent. The ordering of items in the list corresponds to the unmarked word order in a declarative sentence.

**SYNTACTIC FRAME (SF)**, a feature structure with syntactic functions as attributes; values of these attributes are co-indexed with the corresponding items of the subcategorization list.

**UNDERLYING STRUCTURE (US)**, a feature structure with valency (dependency) functions as attributes; values of these attributes are identical with underlying structures within the corresponding items of subcategorization list and syntactic frame. The value of the attribute GOV (governor) is identical with the value of the lexeme attribute of the verb's feature structure.

A parser will establish index links between saturated frame slots and their fillers in the analysis tree. This will provide an easy access to the analysis results at the three levels of description, highlighting the structure of the sentential core.

The three-level complementation paradigms, rather than being stated as full-fledged feature structures, are expressed in a compact form as templates. Other space saving devices, functionally equivalent to lexical rules, will take care of phenomena such as passivization, dative alternation, and there-preposing.

#### 4.4 The Derivation of Valency Information

In an ideal case, a complementation paradigm of the sort described in Section 4.3 should correspond to OALD verb patterns while lexical rules would account for structures listed in Hornby (1974) as variants of the same verb pattern. Although this idea works in the case of the most frequent patterns (VP2A, VP6A), there are many patterns where the relation between pattern and paradigm can be  $1:n$ ,  $n:1$ , or even  $n:n$  ( $n > 1$ ) (see Section 4.2).

The case of  $n$  patterns:  $1$  paradigm reduces the number of paradigms and as such is a welcome situation. The case of  $1:n$  can mean (i) ambiguity for all verbs listed under the pattern (and can possibly be accounted for by lexical rules), (ii) the possibility to subdivide the verbs of this class into  $n$  subclasses, or (iii) a combination of the two. For (i), the derivation of complementation paradigm from a verb pattern will yield a disjunction. For (ii), verbs with different complementation paradigms should be distinguished. Boguraev and Briscoe (1989) used valency codes in LDOCE (*Longman Dictionary of Contemporary English*) to automatically extract the (explicitly unmarked) distinction between Equi and Raising verbs. A similar approach can be used to make this and other distinctions in OALD by taking into account co-occurrences of verb patterns. Our situation is simpler in that we, as yet, make no attempt to treat distinct word senses, and more difficult in that the blurred sense distinctions can have negative effects on any derivation procedure. It remains to be seen whether this will lead to results of sufficient reliability.

The correspondences between OALD patterns and complementation paradigms are stated in the simple cases by rules relating one or more patterns to one or more paradigms – templates. Where possible, frequently co-occurring verb patterns are collapsed into a single paradigm with local disjunction, e.g., VP6D and VP7A for *like* (*swimming / to swim*) give the following template: [ *transitive*, { 2ing | 2inf }, *equi* ], which expands into:

SC < [1] N[nom]<sub>3</sub> , [2] V[*{prespart|inf}*] SC < N<sub>3</sub> > ]<sub>4</sub> > ,  
SF [ SUBJ [1] , OBJ [2] ] ,  
US [ GOV like , ACT [3] , PAT [4] ]

There are two possible strategies for the derivation procedure representing two extremes. The first strategy disregards the actual distribution of verb patterns in the dictionary and attempts to combine results of rule application into a compact and meaningful complementation paradigm. The second strategy starts from a list of all combinations of verb patterns within the dictionary and assigns a rule to every combination.

The first strategy looks like a principled solution, but the application of the rules rewriting a verb pattern code by one or more templates can be a source of unforeseen complexities with the result that too many entries will have to be handled manually. The second strategy is much safer: if there are not too many different combinations of verb patterns it might not be too difficult to state rewriting rules for all of them. However, to make a decision, some statistical analysis is necessary.

CUVOALD lists 5695 verbs with 633 different combinations of verb patterns. 4853 verbs (85.2%) are marked by one of the 56 most frequent combinations (each occurring seven and more times). The first nine most frequent combinations range from a single pattern (6A for transitive verbs) with frequency 1971 to a combination (2A,2C,6A,15B) with frequency 81. At the other end, there are 442 combinations occurring only once, 191 two and more times, 119 three and more times and 77 five and more times.

Another survey was aimed at finding most frequent combinations as proper subsets of full lists of patterns by which the entries are marked. E.g., the combination of three patterns 2A,3A,6A occurs alone in 54 entries, but as a proper subset of a larger combination already in 566 entries.

From the above data it seems that a compromise between the treatment of individual verb patterns and of entire combinations would be most efficient. 119 combinations can already be treated by individual rules quite comfortably while the rest can be composed from results of rules applied independently where more alert supervision is required. It also seems feasible to use the rules for combinations to treat parts of the remaining lists of verb patterns, and perhaps add a few more, selected according to the second statistics.

## 5 Conclusion

Two things are obvious: (i) it is difficult, if not impossible, to find a ready-made source of lexical data immediately usable in a NLP system; and (ii) it is difficult, if not impossible, to build a NLP lexicon of adequate coverage from scratch within a tolerable period of time using limited resources. Our experience confirms that it makes sense to invest some effort in processing available machine-readable dictionaries with the aim of extracting as much lexical data as possible into a reasonably structured pool. However imperfect, such a pool may serve as a starting point on a way towards a much richer lexical data base, built not only from lexicographers' wisdom but also from texts processed to yield a condensation of interesting lexical facts.

## References

- [1] Boguraev B., T. Briscoe, "Utilising the LDOCE Grammar Codes", in B. Boguraev, T. Briscoe, (eds.), *Computational Lexicography for Natural Language Processing*, Longman, London and New York, 1989.
- [2] Dolby, Resnikoff, MacMurray, "A Tape Dictionary for Linguistic Experiments", in *Proceedings of the Fall Joint Computer Conference*, 1963, 419-423.
- [3] Grishman R., M. Kosaka, "Combining Rationalist and Empiricist Approaches to Machine Translation", in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, Canada, 1992, 263-274.
- [4] Hajič J., E. Hajíčová, A. Rosen "Machine Translation Research in Czechoslovakia", *META*, 37 (4), (a special issue ed. by M.C. Cormier and D. Estival), Presses de l'Université de Montréal, 1992, 802-816.
- [5] Hornby, A.S., *Oxford Advanced Learner's Dictionary of Current English*, 3rd edition, Oxford University Press, London, 1974.
- [6] Lehmann H., N. Ott, "Translation Relations and the Combination of Analytical and Statistical Methods in Machine Translation", in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, Canada, 1992, 237-248.
- [7] Mitton, R., "A Partial Dictionary of English in Computer-Usable Form", *Literary and Linguistic Computing*, 1, 1986, 214-215.
- [8] Russell G., A. Ballim, J. Carroll, S. Warwick-Armstrong, "A Practical Approach to Multiple Default Inheritance for Unification-Based Lexicons", Special Issue on Inheritance II, *Computational Linguistics*, 18 (3), 1992, 311-337.
- [9] Sgall P., E. Hajíčová, J. Panevová, *The Meaning of the Sentence in its Semantic and Pragmatic Aspects* (edited by J. Mey), Reidel, Dordrecht / Academia, Praha, 1986.
- [10] TMI-92, *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, Canada, 1992.
- [11] Walker, D.E., "Knowledge Resource Tools for Accessing Large Text Files", in Sergei Nirenburg, (ed.), *Machine Translation: Theoretical and Methodological Issues*, Cambridge University Press, Cambridge, UK, 1987, 247-261.
- [12] Walker, D.E., "Developing Lexical Resources", in *Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary*, University of Waterloo Centre for the New Oxford English Dictionary, Waterloo, Ontario, Canada, 1989, 1-22.

- [13] Walker D.E., R.A. Amsler, "The Use of Machine-Readable Dictionaries in Sub-language Analysis", in R. Grishman, R. Kittredge, (eds.), *Analyzing Language in Restricted Domains*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, 1986, 69-83.
- [14] Walker D.E., A. Zampolli, N. Calzolari, (eds.), Special Issue on the Lexicon, *Computational Linguistics*, 13 (3-4), 1987.

# The IIT Lexical Database: Dream and Reality

Sumali Pin-Ngern Conlon  
University of Mississippi

Joanne Dardaine, Agnes D'Souza, Martha Evens  
Sherwood Haynes, Jong-Sun Kim, Robert Strutz  
Illinois Institute of Technology  
*e-mail: csevens@minna.acc.iit.edu*

## Abstract

This paper describes IITLEX, a lexical database under development at Illinois Institute of Technology. It emphasizes the sources of the data and the organization of the database.

## 1 Introduction

A central interest of many computer scientists, cognitive scientists, and computational linguists has been the construction of a sophisticated lexical source to support all kinds of natural language processing (NLP) (Calzolari, 1988; Byrd et al., 1987; Ahlsweide and Evens, 1988a, McCawley, 1986). Furthermore, if the lexicon is to effectively help computers process language, the necessary information about words and about the world must be included in a highly systematic form that can be easily used by computers.

A major source of information for computational linguistics is the Machine Readable Dictionary (MRD). MRDs contain tremendous amounts of lexical information both implicitly and explicitly. The recent interest among lexicographers in building dictionaries for advanced learners has vastly increased the amount of explicit information available in MRD's, with great benefit to linguists and computer scientists as well as the intended readers (Boguraev, 1993; Sinclair, 1987a, 1987b). However, there is also a great deal of important syntactic and semantic information that is not contained in MRDs. We need to incorporate other types of knowledge beyond that provided in MRDs. A tremendous number of valuable insights achieved by linguists are not being used in NLP.

If the lexicon is going to be of use to computer programs and to people of varying linguistic backgrounds and points of view, all the information required to understand the data must be included in the database itself. Thus, we have included with the Indiana Verb Lists, for example, the sentence contexts devised by Householder's group to determine list membership.

Our intent is to include any kind of information about each entry that may be needed in natural language processing systems. Thus, we include much information that could be described as encyclopedic rather than lexical. We plan to include a superset of

the lexical universe defined by Apresyan et al. (1970). We hope eventually to make it possible to locate the whole semantic field of each entry by threading through the information available in the lexical and semantic relationships that we provide.

To make the lexicon even more useful, the information in it should be refined through comparison to actual usage. This can be acquired using corpus analysis techniques from newspapers, magazines, technical reports, etc., as well as directly through lexical acquisition systems.

The lexicon must also be extensible, allowing such additions as new words and word usages, new linguistic representations, and statistics from analyses of corpora.

This paper discusses the sources of our lexical information (Section 2), describes the kinds of lexical information included (Section 3), explains the organization of our lexical database (Section 4), and outlines the steps we have taken toward implementation (Section 5). The paper concludes with a brief discussion of the ways our ideas have changed through contact with data and with users of the database (Section 6).

## 1.1 How We Got Started

We have always been crazy about dictionaries. Twenty years ago we were arguing with Raoul Smith (1985) about what we would do first, given a good parser. We insisted that the first target should be dictionary entries, so that we could extract still more lexical information to drive the parser further, etc. This argument is painfully naive, but there is a kernel of truth within it. There is still a tremendous amount of information in the definition text in machine readable dictionaries that we have hardly begun to tap. Almost all the work so far has been directed at organizing other, more accessible parts of the lexical entry.

Almost fifteen years ago, thanks to the generosity of the National Science Foundation and the G & C Merriam Company we started to extract information about lexical-semantic relationships from a machine-readable copy of *Webster's Seventh New Collegiate Dictionary* (W7) with the goal of producing a thesaurus for the automatic enrichment of queries in an information retrieval system. The resulting thesarus was effective (Wang et al., 1985).

Our awareness of other work with machine-readable dictionaries was limited to the newsletter for W7 users organized by John Olney until Don Walker and Bob Amsler organized a workshop at SRI in the spring of 1983 involving professional lexicographers and dictionary publishers as well as the computational linguistics community. This meeting was tremendously exciting and informative. As a result Ahlsweide began to parse verb definition text using Sager's Linguistic String Parser (Ahlsweide et al., 1988; Ahlsweide and Evens, 1988a). He eventually parsed eight thousand verb definitions and concluded that while text processing is more effective for obtaining thesaurus information, parsing is essential to the extraction of more complex semantic information (Ahlsweide and Evens, 1988b).

Sumali Pin-Ngern Conlon carried out the initial design of the lexical database and started to fill it with data from *Webster's Seventh New Collegiate Dictionary* (W7) as part of a joint research project with as Edward Fox and J. Terry Nutter to develop lexical data for an information retrieval thesaurus (Nutter et al., 1990). At that point the G & C Merriam Company decided not to allow any further work with W7. It became clear to

us that a lexical resource based on W7 could not be shared even with fellow researchers.

We were rescued by the generosity of Collins Publishers, who gave the first edition of the *Collins English Dictionary* (CED) (Carver, 1979) to the research community and in particular gave us a letter saying that any output from our research could be shared through the Data Collection Initiative with all those that had permission from Collins to use the dictionary tapes themselves. We also spent a lot of time working with the resources that were developed and published by Householder's group at Indiana University and by Jacobson (Conlon et al., 1993; Evens et al., 1991).

The discussions of lexical and textual resources in Walker (1985) have been of overwhelming importance in the development of the field. The international exchange of ideas and resources between publishers, researchers and professional lexicographers would never have occurred without him (Walker et al., 1987). His ideas have been fundamental to our work and, we believe, to much of the work in the field.

## 2 Sources of Information

To provide support for natural language parsing and text generation, the lexicon must contain a wide range of linguistic knowledge. Fig. 1 shows the major sources of information that we use in lexicon construction. These sources are now described in some detail; they include the Indiana Lists (Householder et al., 1964, 1965), the Brandeis Verb List (Grimshaw and Jackendoff, 1985), and *Adverbial Positions in English* (Jacobson, 1964).

### 2.1 The *Collins English Dictionary* (CED)

Data derived from the CED forms the backbone of IITLEX. The CED contains both syntactic and semantic information in structured and unstructured, explicit and implicit forms. Structured, explicit information such as the part of speech designation has been loaded directly into the lexicon. Less structured information, such as the actual definition texts, is being analyzed with the goal of converting it into a more structured, explicit form.

### 2.2 The Indiana Lists

These vocabulary lists were developed by Householder's projects at Indiana University (Bridgeman et al., 1965; Alexander and Kunz, 1964; Householder et al. 1964, 1965). There are four major Indiana word lists corresponding to four major parts of speech: nouns, verbs, adjectives, and adverbs. The verb, noun, and adjective lists describe sentential complement patterns and list words that can appear in these patterns. The adverb list contains placement information and semantic classifications as well.

The following shows the sentence structure for list 2D in the Indiana noun list:

NOUN-PHRASE + PASSIVE VERB-PHRASE +  
THAT + SUBJUNCTIVE SENTENCE  
e.g., The SUGGESTION has been made that he win.

Nouns that can be used in this pattern include *agreement*, *decision*, *law*, *petition*, *recommendation*, and *restriction*.

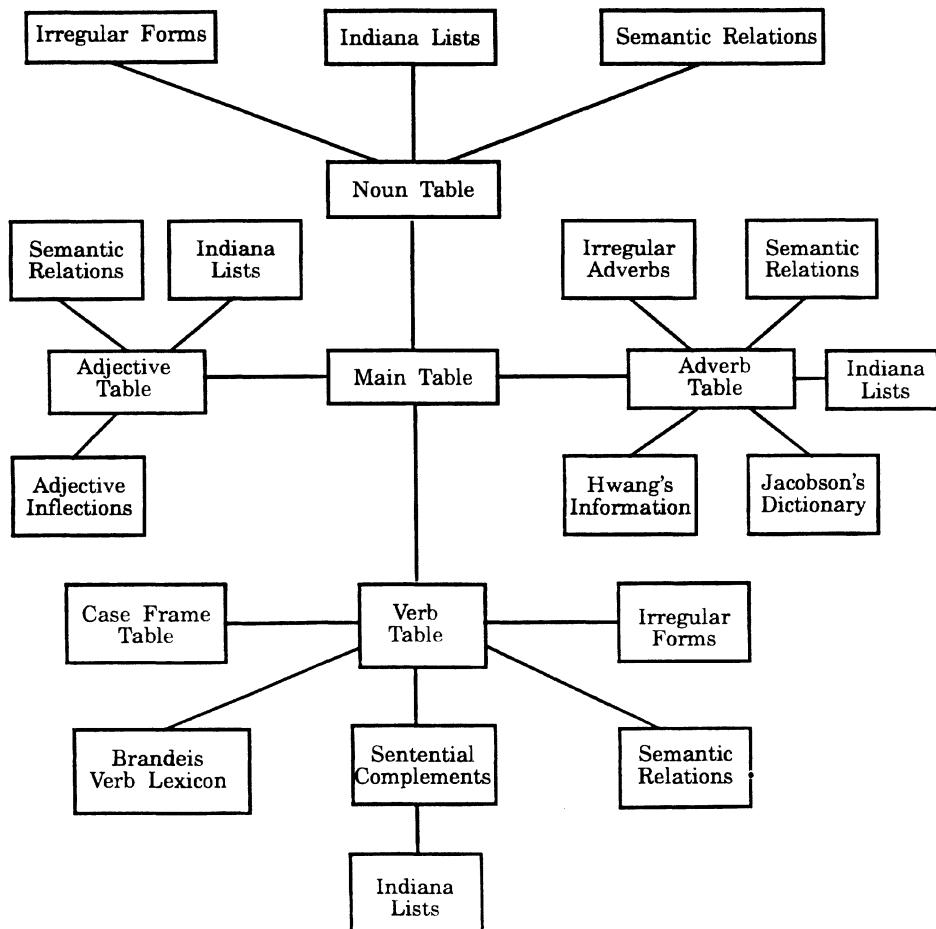


Fig. 1. Organization of the Lexical Database

## 2.3 The Brandeis Verb List

The Brandeis Verb List (Grimshaw and Jackendoff, 1985) contains detailed subcategorization information for about 900 verbs. The possible complements of the word *admire*, for example, are explained as follows:

### ADMIRE VERB

DO POSSING THAT DO-P4-ING-OC DO-P4-NP

where

DO	= Direct Object
POSSING	= Possessive +ING: We admire his helping
THAT	= Tensed complement introduced by "that"
P4	= Preposition "for"
ING	= Gerundive
OC	= Object Control
NP	= Used ONLY for the NP in a prepositional phrase

This entry indicates that *admire* may take an ordinary noun phrase as direct object or a possessive form, etc. If there is any sentential complement then the complementizer will be the possessive form or a noun phrase. (See also Grimshaw, 1990.)

## 2.4 Jacobson's Dictionary

Jacobson (1964) collected adverb placement information by observing adverb occurrences in continuous text. He includes in his dictionary adverb classifications and positions of adverbs as they appear in collected texts. The following shows his position information for the adverb *completely*:

COMPLETELY. Degree; Mod.; 26/14

M (12=63%) M1: "It was the motive of satire which completely dominated his second narrative."

M3: "He had no doubt completely forgotten he ever met his fellow."

M4: "The Duchy of Lancaster, which had never been completely fused with the Crown, was represented by its Chancellor."

E (7=37%) E4: "It is all very well to say she hates it and would go mad completely if she had to go back."

E5: "If she got away with that one it let Focquet out completely."

This entry informs us that *completely* is an adverb of degree. We also learn that out of the sentences surveyed containing *completely*, 63% of the sentences have *completely* in the middle position and 37% in end positions (after the verb). More information about semantic and syntactic properties of adverbs is given in (Jacobson, 1978).

### 3 Lexical Needs of Natural Language Processing Systems

A lexicon should contain information that will support NLP needs. Ideally, every word that could appear in input or generated text or be used by users should be recorded in the lexicon. Walker and Amsler (1986) have shown that this is an unrealistic dream; they found that 40% of the vocabulary of the *New York Times* did not appear in the machine-readable version of *Webster's Seventh New Collegiate Dictionary*. Beyond that, we must try to characterize the needs of information retrieval, language understanding, and text generation systems in terms of language processing. In this section, therefore, we briefly review some of the types of information needed to process texts, as motivation for what should go into the lexicon. Consider, for example, the sentence.

The owners of the cars and the trucks probably thought that they could put them up for sale.

Most human readers would not even notice the potential sources of ambiguity in this sentence, even the most obvious - the question of the referents for the pronouns "they" and "them". To process sentences like this, NLP systems need information about orthography, phrases and idioms, morphology, as well as the syntax and semantics for each word sense (Jensen and Binot, 1987).

Orthographic information involves many spelling-related issues, e.g., "owner", as well as graphemes (symbols like "\$" and "#"), hyphenation, and form variants ("color" vs. "colour").

Phrases and idioms are common expressions consisting of two or more words and having meanings different from the combination of the words constituting the idiom. This type of information is extremely important for NLP systems (Ahlsweide et al., 1988; Markowitz et al., 1988). For example, the language understanding system needs to know that the phrase "put up for sale" is a single lexical entry.

Morphological information includes the way words change their forms in various situations (tense, voice, degree, etc.), e.g., *thought* is the past tense of *think*. This information includes inflectional paradigms and derivational morphology, also.

Information about part of speech is essential for the analysis of sentence structure. In our example, the system needs to know that the words *owners*, *cars*, and *trucks* are nouns, while *could* is an auxiliary verb, and so on.

Further classifications, beyond simple part of speech, can also be very important in NLP applications. Word classifications can help NLP systems to understand and disambiguate sentences in language understanding systems, assist word selection in text generation systems, select pronouns effectively, and so on. To process this sentence, the system needs to know that *think* can take a sentential complement introduced by *that*. We need to have information about how to combine verbs with their arguments. Thus, we have to include, for example, the fact that *think* needs a human or animate agent, that it does not require an instrument, and that it takes sentential complements. How does the word *owners* from our example sentence fit into various dimensions along which nouns are classified?

1. Regular vs. irregular nouns: *owner* is a regular noun, meaning that the plural form can be built by adding the suffix *s* to the root *owner* yielding *owners*.

2. Countable vs. uncountable (mass) nouns: *owner* is a countable noun. Thus, for example, its singular form must ordinarily have an article *an* or *the*.

3. Human, animate, or inanimate nouns: *owner* is a human noun. This tells the system that an *owner* can serve as subject to verbs that require human subjects. For example, *owners* can *think*. This is useful for language understanding and knowledge representation purposes.

Similar word classifications exist for verbs, adjectives, and adverbs, as well as nouns. As an illustration, adverbs can be classified as conjunctive or linking adverbs, sentence adverbs, manner adverbs, etc. (Conlon and Evens, 1992). This classification can be very useful in parsing, language understanding, and text generation.

For example, the fact that *probably* in the example sentence is a sentence modifying adverb has implications for the parser trying to figure out where it belongs in a parse tree, since its sentence position would suggest that it is part of the verb phrase. The classification is therefore necessary for proper parsing. Similarly, the fact that *probably* should link to an S-node tells text generation systems that it could not be placed between a verb and a particle, e.g., between *put* and *up* in our example sentence. Finally, if a text generation system is to avoid generating ambiguous sentences, it also might be designed so as not to put *probably* at the end of sentences like our example. In the end position, *probably* could attach to either of two S-nodes, and so, be interpreted as in either “probably thought” or “probably could put up for sale”.

Lexical semantic relations such as ISA and PART-WHOLE are important since they allow the system to find the connections between word senses. This type of information is helpful in both parsing and text generation (Miller et al., 1991; Morris and Hirst, 1991; Lee and Evens, 1992).

Information about synonymy can help systems to know when different words refer to the same object or event. This relation is represented using the lexical semantic relation SYN (e.g., *buy* SYN *purchase*).

The antonymy relation provides information that seems to be crucial in making lexical choices in surface generation. Justeson and Katz (1991) have provided strong evidence for the hypothesis that pairs of antonymous adjectives like *good* and *bad* co-occur with very high frequency.

Information about selection restrictions or preferences helps parsers and language understanding systems select appropriate word senses in cases of lexical ambiguity. By using selectional restrictions (preferences) the system will know that *the trucks* cannot *think* (because *truck* is an inanimate noun). Thus, only *the owners* could have *thought*, not both the *the owners* and *the trucks*. Selectional restrictions eliminate some possible parses for the example sentence.

Now that we have briefly reviewed some of the types of orthographic, morphological, syntactic, and semantic information that are needed in the lexicon, we turn to a discussion of the organization of this information in the lexicon.

## 4 Organization of the Lexicon

Entries are separated by parts of speech: nouns, verbs, adjectives, and adverbs. IITLEX has many tables corresponding to each part of speech. Fig. 2 shows the overall structure of IITLEX.

The main table contains word entries with homograph number, sense number, and parts of speech assigned by the CED, as well as words and phrases from other sources. All the entries have a unique identification number assigned by a program. This identification number is used in look entries up in different tables and in combining tables in different ways. The main table serves as a kind of index to IITLEX; it indicates what other tables contain this entry.

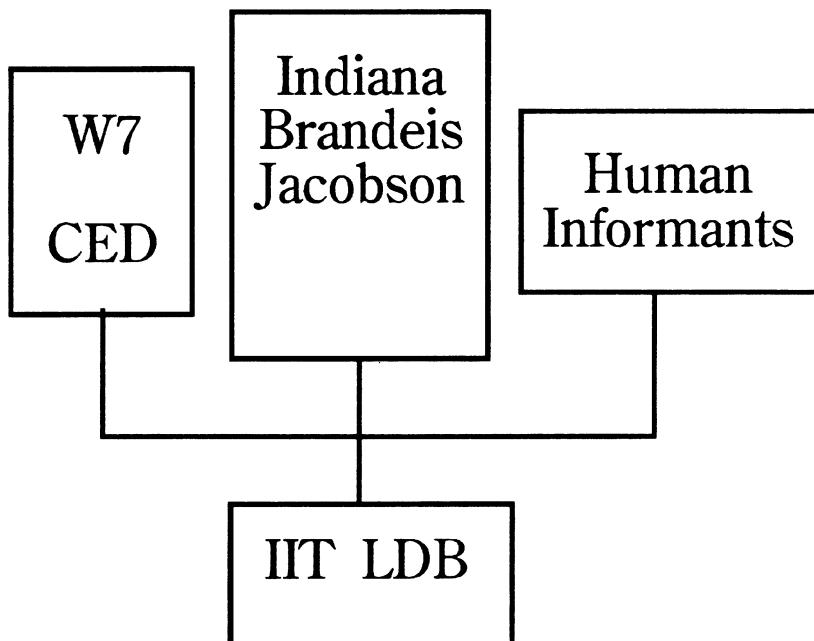


Fig. 2. Sources of Information for the Lexical Database

## 4.1 Nouns

In the noun lexicon, we specify whether a noun is regular or irregular, singular or plural, abstract or concrete, common, proper, or collective, count or mass, abstract or concrete, and human, animate or inanimate. In addition, we need to indicate noun genders to help NLP systems interpret references correctly and generate appropriate pronouns. More information about our noun lexicon can be found in Conlon and Evens (1994).

## 4.2 Proper Nouns

Before we began to study proper nouns we assumed that they were just like other nouns and belonged in the noun table. Any attempt to process newspaper text reveals that they present many novel problems. Indeed, proper nouns are a major factor in the 40% of the vocabulary of the *New York Times* that Walker and Amsler (1986) could not find in the MRD's they consulted. One advantage of working with the CED is that it contains regular entries for many proper nouns instead of excluding them or isolating them in a gazetteer or a biographical section. Hoang et al. (1993) have made a preliminary study of the entries for proper nouns in the CED. They propose separate tables for people, organizations, naturally existing geographical locations (such as mountains and rivers), man-made places (such as cities and countries), and times.

Jong-Sun Kim has begun a systematic effort to understand the use of proper nouns in the *Wall Street Journal* and to derive information for lexical entries from it. When the *Wall Street Journal* introduces a proper noun that it does not expect the reader to recognize, it provides information about that person, place or organization that is remarkably similar to the kind of material provided by the CED. For people it gives an age, title, nationality, and either a profession or an organization and a position within that organization. For an organization it gives a name, its acronym if one is in common use, and the major focus or foci of that organization. Places are located roughly in terms of country or area and also in relation to a place it assumes everyone will recognize. The main table for proper nouns (Table 1) indicates what class each noun belongs to. We are constructing separate tables for the different categories of proper nouns.

Name	Category	Type	DefArt
Bass Strait	wildgeog	channel	n
Canada	humangeog	country	n
Fortune	document	magazine	n
Harvard University	organiz	college	n
IBM	organiz	computer	n
Kent	product	cigarette	n
King of Thailand	person	govern	y
Ottawa	humangeog	capital	n
State Department	organiz	govern	y
Vinken	person	corporate	n

Table 1. A Few Entries from the Main Table of Proper Nouns

## 4.3 Verbs

Verbs seem to be both syntactically and semantically more complicated than other parts of speech. We classify them as regular/irregular, dynamic/stative, and transitive/intransitive. These classifications can be found either directly from the MRDs or by simple analysis of MRD definitions. Some complex syntactic information useful to NLP systems needs to be found by hand such as whether a verb takes a sentential complement or not, whether it can be put in passive form or not. Some of these entries are still incomplete.

Case frames for verbs are important in parsing and generation (Hirst, 1986). Our case frame tables contain information relating the syntactic and semantic roles of the arguments of the verb and the selectional restrictions those arguments must satisfy.

## 4.4 Case Frames

The Case Frame Table (Dardaine, 1992) is being developed in an attempt to provide users of the lexicon with some of the information they need about the arguments of verbs. What kind of information do users of the lexicon need about verb arguments? They need to know what argument slots are present for a given verb sense; then, for each slot they need to know what the associated case or semantic role may be, how that role is expressed syntactically, what selectional restrictions operate on that slot, and also whether that slot must be filled. The names of the cases we are using are taken from (Allen, 1987), based on Fillmore (1968). For the benefit of human users we also include an example sentence for each sense/case frame combination. The first customer for this information is an intelligent tutoring system for cardiovascular physiology, so the sentences included in the examples below use sentences from actual tutoring sessions. (The tables also include identification numbers for word senses and case frames that are not shown because they are not meaningful for the reader.)

Verb	Syn_role	Case	Occurrence	Select-restrict
accomplish	subject object	agent theme	obligatory obligatory	human/reflex action

e.g. Think about what the reflex is attempting to accomplish and I think you can tell what CO will do. (Circsim K23-28)

act	subject	agent	obligatory	reflex
-----	---------	-------	------------	--------

e.g. Yes, since TPR is determined by sympathetic activity, there can't be any change before the reflex acts (which it doesn't do in DR). (Circsim K23-28)

return	subject object to-NP	agent theme cotheme	obligatory optional obligatory	parameter/event parameter value/parameter
--------	----------------------------	---------------------------	--------------------------------------	---

e.g. This would tend to return VR to P.

From an abstract point of view, the Case Frame Table is a complex data structure that contains in it syntactic and semantic, as well as pragmatic information. This information is vitally important in parsing and text generation.

The design of IITLEX's Case Frame Table differs from other proposed systems. Rather than a comparative study to show the frequency of a particular case for a group of words as done in (Calzolari, 1988), the IITLEX Case Frame Table is, by design, verbose and redundant. The decision by the IIT Lexicography Group to incorporate as much information as possible, regardless of redundancy, is reflected in the design of this table.

The Case Frame Table information includes the following columns:

Verb – These verbs are contained in IITLEX\_Words.

Word\_Id – Unique identifier given to each word in the main table, IITLEX\_Words, and used as a means of accessing other tables in the lexicon.

Case\_Id – Unique identifier given to each case frame entry for a given word. The existence of multiple mappings, discussed elsewhere, warranted the inclusion of this field.

Syn\_Role – The syntactic role of the verb argument. Includes Subject, Object, and Adjunct.

Case – Semantic roles of verb arguments as defined in (Allen, 1987). Includes Agent and Co-Agent, Theme and Co-Theme, Beneficiary, Experiencer, at-Poss, to-Poss and from-Poss, at-Loc, to-Loc and from-Loc, at-Value, to-Value and from-Value, and, at-Time, to-Time and from-Time.

Occurrence – One of the three different values used to identify valid syntactic sentence constructions. These values include *Obligatory* for necessary arguments, *Optional* for arguments that can be omitted, and *Elliptical*, for arguments that can be omitted if the context is known.

Select\_Restrict – The pragmatic restrictions on the arguments of a verb. This is based on a noun taxonomy by Agnes D'Souza. Restrictions can be general, such as the restriction on the object of the verb *read*, namely, *document*, which encompasses all printed material such as books, journals, magazines, etc., or as specific as the restriction on the object of the one sense of the verb *lay*, namely, *egg*.

Example – An actual sentence, taken from a variety of sources, including current literature, newspapers, dictionaries, etc., that relates the case frame structure to its appropriate use in a sentence.

As in most design projects, the design has changed many times. Dardaine has done extensive manual analysis over the last two years, and has discovered a number of significant problems. In solving them, she has made a number of modifications to the design proposed by Pin-Ngern (1990). Dardaine has since developed frames for about 120 verbs (200 senses) from the cardiovascular domain and about 800 more (about 1500 senses) that appear in COBUILD and that begin with the letters A through D.

At the beginning of the work on the Case Frame Table we had naively hoped that one CED verb sense would have one Case Frame. As we mentioned in the discussion of Case\_Id in the Case Frame Table, the inclusion of the attribute, Case\_Id, is a direct response to the fact that verbs display a multi-mapping behavior between senses and their corresponding case frame structures. Some verbs, generally those with only one

sense, such as *pantomime*, have a 1-1 mapping, meaning that the verb has one distinct corresponding entry in the Case Frame Table. Others, generally those in which the senses are so closely related that the case frame entries can be combined to form one single entry, exhibit a M-1 mapping, meaning that many senses are mapped into one case frame entry. An example of this is the verb, *mutter*. Yet others, especially verbs that take a variety of prepositions in their adjunct definition, exhibit a one-to-many mapping, meaning that a single sense may be mapped into several case frame entries. One example in point is the verb, *arrive*. We have three separate frames for this verb.

arrive	subject	theme	obligatory	object
	adjunct	to-loc	optional	location

e.g. I arrived safely. I arrived home safely. News of the attack arrived here yesterday.

arrive	subject	theme	obligatory	object
	adjunct	from-loc	optional	location

e.g. I arrived from Trinidad last week.

arrive	subject	experiencer	obligatory	human
	at_np	result	obligatory	idea

e.g. Do you know that we arrived at that same conclusion yesterday?

Another event that had a profound effect on the structure of the Case Frame Table was the realization that the syntactic roles should not be limited to subject and object as proposed in the original design (Pin-Ngern, 1990). The rationalization for this modification was the fact that other verb arguments had associated with them semantic and pragmatic information, and if the IITLEX system was to be as complete and robust as possible, that information should be represented. This led to the inclusion of other verb arguments, such as Obj\_Reflex for reflexive objects, Obj\_Indirect for indirect objects, Exc for exclamations, Nom for predicate noun phrases, and INFQ for tensed indirect questions, among others.

The determination of which arguments to include and how much information should be included poses a problem, particularly when it comes to sentential complements. This problem involves both theoretical and practical concerns. So far we have separated the information in the case frame table from the information about sentential complements available in the Indiana Verb Lists, which tell us about complementizers, aspects, tenses, and third person subjects for about 1700 verbs. We are concerned about whether separation is valid or not. Does the nature of the sentential complement interact with the selectional restrictions and case roles of the other arguments? Does it do so often enough so that the case frame table needs to include the information in the Indiana tables? We do not know.

The practical side of this question is far from minor. On one hand, we want a system that is as complete as possible. On the other, we also want to generate as much

information semi-automatically as possible. Our sources of information for generating this information are limited and involve a finite number of verbs. The most detailed source is the sub-categorization information produced for 900 verbs by Grimshaw and Jackendoff (1985). We intend to use the Brandeis Verb Lexicon to generate case frame information.

The Brandeis Verb Lexicon tells us that the verb, *acknowledge*, takes a tensed indirect question introduced by *whether*. This information is available in a somewhat different form in the Indiana Verb Lists. Should we put this information in the case table or leave it in the tables representing the data from Indiana? This cannot be done properly until the Indiana information is annotated with the CED sense numbers.

The research on case frames is closely related to our work on Noun Taxonomy, since the selectional restriction field of the Case Frame Table is usually filled by nouns and noun phrases. The implication is that if X is the filler of the selectional restriction slot for the subject of a given verb sense, then any noun that appears below X in the Noun Taxonomy can serve as subject for that verb. (We hope that this is true often enough to be useful.) Thus, the Case Frame Table assumes that all verb arguments can be classified. This is a long and complex process which is very much outside the scope of the case analyst. However, we hope that, by working closely with Agnes D'Souza, who is developing the noun taxonomy, the case analyst can develop a cohesive, carefully developed, and integrated system in which the Noun Taxonomy can be used to make inferences from the selectional restrictions of the Case Frame Table.

Sometimes during manual analysis we realize that we have developed the same case frame for two different senses. We started to coalesce case frames that were otherwise similar but had different selectional restrictions but then realized that we were making a mistake. In fact, the pragmatic selectional restrictions of the object may not be compatible with those of the other arguments. Consider the verb *agree*, which has the same values in all fields except the selectional restriction. One might think that for the various senses, one can combine the restrictions of the arguments of the OBJECTs of all senses, the SUBJECTs for all senses, and the WITH\_NPs for all senses. However, doing this would mean that the sentence

The climate agreed with the proposal

would be valid. This is clearly a pragmatically incorrect sentence. Levin (1987) and Atkins (1993) suggest still other factors that we should be considering.

## 4.5 Adjectives

Very little work in computational linguistic research has focused on modifiers such as adjectives and adverbs. There are many classes of adjectives that modify nouns, and these classifications have interesting implications for adjective position in a sentence. Quirk et al. (1985) separate noun modifiers into four groups, which are:

1. Zone 1: Precentral – this group includes peripheral, nongradable adjectives, such as *certain*, *definite*, *sheer*, *complete*, and *slight*.
2. Zone 2: Central – this includes adjectives that satisfy the four major criteria of adjective status, which are

- 2.1 They can take on an attributive function (i.e., can modify a noun and appear between the determiner and the head noun, as in *an ugly painting*, *a round table*),
  - 2.2 They can take on an predicative function (can function as a subject complement, e.g., *the painting is ugly* or an object complement, e.g., *he thought the painting ugly*),
  - 2.3 They can be premodified by the intensifier *very* as in *the cars are very clean*,
  - 2.4 They can take comparative and superlative forms, such as *happier*, *largest*, *more complicated*, etc.
3. Zone 3: Postcentral – this includes participles (e.g., *retired* and *sleeping*) and color adjectives (*red*, *black*, and *white*).
4. Zone 4: Prehead – this zone includes the least adjetival and the most nominal items, such as *Australian*, *Middle Eastern*, and *American*. (Quirk et al. 1985:402-403,4).

In building noun modifiers, the lower zones must appear before the higher zones (such as *a clean red middle eastern rug* but not *a red clean middle eastern rug*). It will be very useful to include these adjective classifications in the lexicon so that the system can determine where they can appear as modifiers.

Sowa (1988) argues that the lexicon needs information about arguments of adjectives like that for verbs. There are many parallels between verbs and adjectives. Adjectives may also be either dynamic or stative like verbs. Some adjectives support sentential complements like verbs. More generally, a given adjective sense typically applies to some nouns and not to others. In other words, we should be storing selectional restrictions for adjectives. At this point we have only the Householder et al. (1964) Indiana List information about adjectives (which describes sentential complements).

Text generation systems also require information about adjective position. Some adjectives cannot appear in attributive position: For example, I may say, far too often, *my class is asleep*, but I cannot refer to *my asleep class*. Some adjectives cannot appear in predicate position (Levi, 1973): *an electrical engineer* but not *the engineer is electrical*. A few adjectives like *galore* prefer to appear after the noun they modify. The CED contains some information about adjective position; we are now working on coding that information and supplementing it.

## 4.6 Lexical Semantic Relations

The lexical database contains a list of lexical semantic relations and their properties. That is, it indicates for each relation whether it is reflexive, symmetric, transitive, or one-to-one. Another table contains word-relation-word triples, and is so large that it is actually stored in three parts. The first part contains taxonomic information; the second part contains information about synonymy; the third contains the triples for all other relationships. This organization is purely a matter of convenience; we have more information about taxonomy and synonymy because they are more often explicitly indicated in CED. We hope to expand the information available about other relationships.

## 5 The Current Implementation of the Lexical Database

The implementation of the lexical database has involved three major steps. The first step was the analysis of the information in a machine-readable version of the *Collins English Dictionary* or CED. The second step involved populating the flat files that underlie the

tables in the lexical database. In the process our ideas about the database structure have changed several times. The third step, still ongoing, is the effort to provide appropriate tools to access the database, to produce datasets for users, and to update the database.

The problems of managing and merging some of our other lexical resources are mentioned briefly in the discussion below. The Indiana Lists (Householder et al., 1964, 1965) have been rendered into machine-readable form, along with the adverb dictionary from (Jacobson, 1964). The Brandeis Verb Lexicon (Grimshaw and Jackendoff, 1985) is stored as a separate file; it has also served as a major resource in the construction of the Case-Frame Table. The major problem has been matching word senses to those in the CED. This work has been completed for the Adverb Lexicon but remains to be done for some of the other data.

The goal was to make the system a useful source of subsets or sublanguage dictionaries of all kinds, and, at the same time, to make it easy to add new information. As we moved from the design phase to implementation of the lexical database, several issues influenced our thinking: (a) the lexicon involves many complex structures; (b) the lexicon is an ongoing work-in-progress with more information being added all the time; (c) as research continues, the basic structure of the lexicon may change. We have attempted to address these concerns and still create a robust lexical resource that can provide comprehensive data for text generators, parsers, and information retrieval systems at the level of detail and complexity that the user wants. We needed data representations and mechanisms to access the data that are meaningful both to computer programs and to humans.

## 5.1 Parsing the *Collins English Dictionary*

The backbone of our lexical database was produced by Robert Strutz by parsing a machine readable version of the first edition of the CED. The CED was chosen because it enabled us to provide a large and consistent base of lexical information that can be made available for a wide range of research purposes. With the permission of the publisher, we obtained a tape from the Data Collection Initiative designed to drive a commercial typesetting machine.

There are three places in the source file where the data has been corrupted. Some of the text is duplicated and some is missing. A published version of the CED was consulted to resolve these problems. These are the only manual modifications made to the source file.

The tape still has the tags used to change the weight and type of fonts, identify different sizes and types of hyphens, spaces, and punctuation marks, and other special characters. These tags represent information that could be lost if the file was parsed directly into ASCII without also decoding and representing the information in the tags. Much of this knowledge is taken for granted by users of dictionaries. The human user understands the content of the text and ignores many inconsistencies and errors in font type and weight, mismatched parentheses, brackets, and braces, and so on. In addition, people see the font changes, bold elongated hyphens, and italics, and infer that the highlighted words and phrases have a special meaning. For the material to be consistently and correctly accessed by machines this implied knowledge must be explicitly represented. By referencing a published copy of the CED, we were able to

resolve many conflicts in the source text.

The parser is organized as a series of nine passes over the data; each pass is carried out by a separate program. Each program in the series takes the output file from the previous program as its input and produces a new output file. There are six Icon programs and three C programs. Some of these are very simple and do one or two simple tasks; some are quite complex. As we found and translated a new pattern, we needed only to change the appropriate program file and rerun those programs that were dependent on that program or its output. We wrote a series of UNIX style make files and shell scripts to automate the maintenance of the tagged intermediate files.

Within these programs some functions deal with special fonts; some functions deal with specific strings. For example, the function `do_another` processes the strings “another name for”, “another term for”, and “another word for”. The typesetting in the source file is not completely consistent. For example, a few “another” strings have imbedded font changes. Instead of just disregarding the font changes, it is necessary to check each instance that does not exactly match the defining formula to ensure that there is no implied knowledge that would be lost.

The first edition of the CED contains several classes of typographical errors. Most of these are cosmetic in nature, for example, mismatched parentheses and brackets. A casual human user would not notice this kind of error. A computer program that tries to identify structures in the text by matching symbols would produce inaccurate results without producing an error message. The left and right parentheses and brackets do exist but are not properly nested. We wanted to leave the original data in the exact form in which we received it, but we wanted to repair these errors as early in the process as possible. At first, we produced a UNIX style context difference file between the original and the “corrected” version, but this approach did not work as well as expected. The source file is a stream file over 30 megabytes long and the difference program produced a file that is almost as large as the source file. Instead, we wrote an Icon program to search out and repair these problems. This program, among other things, determines where missing parentheses and brackets should be placed. For example:

```
#Hsham#+an #5(#!#s@am@n) #6n.@n#1$D. #5a priest of  
shamanism. @n#1$D. #5a medicine man of a similar  
religion, esp. among certain tribes of North  
American Indians.@n[C17: from Russian #6shaman,  
#5from Tungusian #6s*uaman, #5from Pali #6samana  
#5Buddhist monk, ultimately from Sanskrit #6*'srama  
#5religious exercise.@m?-#1sha#+man#+ic  
#5(#s@@#!m@an$Ik)@f#6adj.
```

This entry is missing a right bracket. But the right bracket should follow the word *exercise* in the last line. If the bracket is not correctly placed, the adjective form of the entry will be lost.

As we studied the CED and found implicit information that we had missed before, we changed the Icon programs to identify the data. We determined a small, well-defined set of tags for use in the parser output. Sherwood Haynes (1993) wrote a program to compare the result of our parse with the parse produced by Lloyd Nakutani and distributed by the the DCI. This process has helped to identify a number of errors.

The CED parser output is tagged so that each block is delimited by **BEGIN\_BLOCK:** and **END\_BLOCK:** tags. In the CED, itself, a block is delimited by a blank line (white space). A block may contain many senses but only one homograph number. Homograph numbers are tagged as **HOMO: n**, where n is a number. If a block in the CED does not have an explicit homograph number, then homograph records are not created in the parsed file. The homograph number in these cases is assumed to be 0 for purposes of completeness and cross-referencing with other sources. Sense numbers are tagged as **SENSE: nx**, where n is a number and x is a letter. In the CED sense numbers may be a number alone, (1), a number followed by a letter (1a) or a letter alone (a). If an entry has subsense letters but no sense numbers, the parsing program inserts a 0. As a result, all sense numbers in the parsed file begin with a number and may be followed by a letter.

Next, the head words or main entries are broken out. These are tagged as **ENTRY:** followed by the word or phrase. We tagged parts of speech as **POS:** followed by the category name. Different parts of speech have different attributes, for example nouns have number and are tagged as **NUM: plural** or **NUM: singular**. The CED also describes nouns as **mostly singular** or **usually plural**. This poses problems when inserting the entries into the database. The constraint information on usage could be most welcome to a linguist, but could be difficult for a program to understand. It needs to be coded. Sometimes the number information is associated with a single sense of the entry, sometimes it is associated with all senses of the entry. This again is implicit in the parsed CED and is resolved when the data is inserted into the database. Other structures within a block are tagged in similar ways.

In the CED run-ons may appear throughout a block. (“Run-On” is the term that dictionary publishers use for vocabulary items that are introduced and sometimes defined within the entry for another word. The CED contains thousands of run-ons; most adverbs ending in *ly* do not have entries of their own; instead they appear as run-ons in the entry of the adjective from which they are derived.) Run-ons are identified in the CED by a hyphen followed by a suffix. The CED parser reconstructs the derived word and tags it as an entry. The database parser creates records in the database identifying this derived word and its root.

There are many defining formulas in the CED. Phrases like “Compare <word>” indicate a link. The phrases “term for”, “symbol for”, and “name for” combine information about register with synonym references. There is much other information in the CED that is hidden still more deeply; we hope to identify more knowledge as the process continues.

There still may exist some strings that are not tagged, but the development of the parser has stopped, for now, although some minor changes are being made. For example, there remain about 15 blocks that have unresolved run-ons. We expect to find further problems as we continue to work with the data.

One typical request for a subset of our lexicon comes from a group at Chicago Medical School, whose long range goal is to produce a consistent instrument for the diagnosis of schizophrenia (Rapp et al., 1990). They need an affective lexicon. Agnes D’Souza has been working on the production of this subset. She began by determining the relevant subset of the noun taxonomy that she has been developing for IITLEX. Now she is looking for other relationships between these words and also for links to the adjectives (*joyous, miserable*), verbs (*want, hate*) and phrases (*be on cloud nine*) that

make up such an important part of the affective vocabulary.

## 5.2 Populating the Tables with Data from the CED

Our first design for the lexical database (Pin-Ngern, 1990) was based on an Oracle Relational Database Management system. While we are still convinced that this plan was a valid one and we plan to use this system for acquiring and updating lexical data, we are at this point storing the database in large flat files. While this approach does not provide us with a comfortable assurance of the security and integrity of the database, it means that we can provide access to the database for people who do not have the funds or the technical experience for a relational database management system.

These flat files are constructed in layers. The core or kernel of the database is composed of three tables. The referential integrity of the data is maintained by a series of unique identifiers that are incremented as records are created. This design allows a simple program to populate the tables without human intervention.

The main kernel table, IITLEX\_Words, has two columns, one for the word or phrase, and the other for the word identifier. This number is unique for a given word in the database. Regardless how many times the word or phrase occurs in the database, or how many senses or parts of speech it has, or how many sources it has been derived from, it has only one identifier and only one record in this table.

Throughout the database, the case of words and phrases is always maintained. For example, proper nouns are stored with the initial character upper-case. If the same word exists in the database with different case, it is treated as a different words. Given the choice between missing information and having duplicate information, we always choose to have duplicate data.

The second table named IITLEXEntries, has a column for the word\_id, which is a link to the first table, a column for the tag and the tag's value, a column to uniquely identify the record, and a column to indicate the record's parent.

Tags are used to identify word attributes like sense number, homograph number and gender. Also flagged are relationships such as: "shortened to", "short for", "technical name for", "abbreviation for" and "another name for". Most of these tag names were influenced by the CED labeling conventions. As the lexicon grows, we may want to replace some of them with more general or more meaningful names. These changes will be implemented as the database is created, leaving the parsed dictionary file intact. We now tag over 50 word and relation characteristics. With this design, attributes can be added just by adding tags. It is not necessary to add columns and change the structure of the base tables every time a new attribute is identified.

The third table, IITLEX\_Text, has six columns: word\_id, parent\_id, entry\_id, text\_id, tag, and value. This table contains the definitions, the references, and the usage examples. These are very long, wordy strings that need to be part of the lexicon, but are not specific enough to be machine parsed and tagged. The text in this table is designed to be used by experts to determine the accuracy of the data in the lexicon and to extend entries to fill in any voids in the lexicon.

These three tables are sufficient to represent the data in the CED. Some data, by its very nature, will not easily follow this design. Dardaine's (1992) case frame data is a good example. The Indiana Lists (Householder et al., 1965) is another good example. Also,

this design is very machine friendly, but we want to provide user friendly, interactive access to the data. Both these issues can be addressed with the same solution.

It is the enforced uniqueness that makes duplicate data manageable. At any time, a simple report can be executed against the database producing a listing of all possible exceptions, not just duplicate data, that can be evaluated by linguists or other experts to determine if corrective action is required.

The parse of the CED is intended, as far as possible, to be a true representation of the source file. Any interpretation of the data is left for the program that builds the database. In tagging the CED we wanted to maintain as much flexibility as possible. The output of the CED parser is intended to drive the procedures that populate the database. It is the database from which the subsets will be produced. The original version of the CED does not provide the flexibility and auditability that we need. We must be able to modify, correct, update, and add to our lexicon. A subset generated from a newer version of the lexicon should include all the improvements and extensions that have been added to the lexicon. The goal is a functional, comprehensive lexical resource that will be a dynamic expanding source of lexical data for many forms of research. The parsing of the CED and other machine readable sources of lexical information will produce a large body of relatively consistent data. The database will maintain the origin of all data, whether from published dictionaries or other lexical research. The data that populate this database are a means to that end. Other sources of data are being used to round out the lexicon. Any lexical issues will only be addressed by modifying the database through a series of "tools" that will track who changed what and when the change was made. This level of auditability is not possible if the flat data files are edited directly, or if code is modified on the fly.

The main disadvantage of the flat file format described above is that sorting the file by anything other than one of the unique identifiers is no real help. And the unique identifiers are not meaningful to users. Program access and access from specially designed screens is no problem because the user is distanced from the nuts and bolts of the system. Any direct user access of the flat files means that the user must deal with the nuts and bolts. Direct access also requires that the users enter information about the update source, and give the time and the user identification information. The programs and screens take care of these tasks automatically for the user.

If the relationships in the lexicon were mapped, the resulting structure would be a network, a very big unmanageable network. Every time the lexicon was updated, the network might have to be updated or rebuilt. If the lexicon were being updated often, it would be unrealistic to keep the network current. Also, it is likely that only part of the network would be applicable to any one query. The rest of the network would create overhead for the rest of the system. All of this discussion assumes that the user has a well-defined query, which is not likely. Since the data that composes the lexicon are expressed in natural language, a user might not know how to formulate a query until acceptable results are produced. The user specification guides the construction of the network and defines the format of the output. Both procedures, creating the network and writing the output, use pattern matching and, maybe, different constraints.

Suppose we turn from the back end of the system to the front end. There are two realistic choices: point and shoot menuing, or a query language. The advantage of the menuing system is that there can be no surprises. This type of interface is relatively

straight forward to implement. The disadvantage is lack of flexibility. The only queries that can be executed are those that have been anticipated.

### 5.3 Tools for Accessing and Managing the Lexical Database

Robert Strutz is designing and building a collection of tools with two main functions, to create subsets of the lexicon for use in other projects and to update the lexicon with data from other sources. Planning for tools has gone hand in hand with the design of the database, since we expect that all access to the database will be controlled by these tools.

Requests already received for data from the database suggest that we will need to be able to produce a wide range of subsets. A very simple subset would be created to satisfy a query for all nouns in the lexicon. A more frequent query might ask for information about senses of a list of words from a given sublanguage. We have also received requests for thesauri.

Complex queries raise some very interesting issues. These issues include: What data should be explicitly defined in the lexicon? What data should be calculated and saved in files for future reference? What data should be calculated on demand? Glenn Mayer wrote a program for an earlier version of the lexical database that calculates regular noun and verb forms on demand rather than storing them. We have considered using one of several new morphology programs in this way. At the present, however, we are storing all forms. We hope, also, that some of the users of subsets of the database will give information back to us in augmented form. Tools are also needed to insert this information back into the database without losing the original data.

To facilitate access along many dimension of meaning, we have built a main table in which any word or idiom can be found alphabetically, as well as tables for each part of speech, a special table for words that take sentential complements, etc. Given a root, IITLEX can supply all the forms based on that root. Given a complex form, IITLEX can find the root. IITLEX also contains tables of lexical and semantic relations and relationships, to make it possible to find terms related to any given word in a variety of different ways.

Byrd (1989) and Boguraev and Levin (1990) argue that one of the most crucial properties for a lexical database is open-endedness. In order to allow for future increases in the amount and quality of information, the lexicon should be designed to be easily extendable. For example, it should allow informants or even automatic lexical information systems to add or to modify information in the lexical database based on processing of explicit linguistic information or information contained implicitly in corpora, for example. We have written SQL forms to collect information from individual native speakers of English to facilitate the incorporation of new lexical entries and other information to our database (Evens et al., 1989).

We also want to be able to accept new data from users as they extend the subsets we give them and integrate that new data into the database. We have not yet determined how to carry out this part of the process without matching word senses by hand. We are hoping that current research by Ahlsweide and Lorand (1993) and by Véronis and Ide (1990) will provide semiautomatic methods for solving this problem. Once the data are tagged with proper word sense information, an intelligent tool is needed to add that new

data to the lexicon.

## 6 Conclusions

This paper describes the contents and the organization of our lexical database. Our lexicon is now available in two forms, flat files and Oracle database management systems. The flat file version allows users without access to Oracle to use our lexical database using their own programs. The Oracle version makes it easier to link different types of information about each word easily by joining tables that share common values.

Our ideas about how to store the data have changed radically over the last few years as our ideas of how our lexicon would be used have changed. We originally viewed our lexicon as being directly accessed by other people's programs; now we see our lexical database as a lexical resource from which we will produce sublanguage lexicons for use by other people's programs. We initially planned to store all of the data all of the time in a relational database management system. Now we are storing the data in large flat files, but using a relational database management system when human beings update the data or want to query it, because of the ability it gives to provide and acquire data through a forms interface. We are still admirers of the Oracle systems but we are planning to convert to a database that is available free to researchers.

Our basic goal has been to make as much information as possible explicit and accessible. The most important word classification is part of speech (noun, verb, adjective, adverb, preposition, etc.). In addition, words that belong to the same part of speech but have different characteristics are treated separately. This is especially important because different types of information are relevant for different classes and subclasses of words. Each word sense is treated as a separate entity, since it may belong to a different part of speech than other meanings conveyed by the same string or have other syntactic and semantic characteristics of its own.

Many lexical entries consist of more than one word. Often their meanings are different from the combination of the meanings of the words from which they are built. This type of lexical item will not be recognized by parsers and language understanding systems or used properly by text generation systems if IITLEX does not list the whole phrase together. (Ahlsweide et al., 1988) and (Markowitz et al., 1988) describe the phrases in our lexicon in more detail.

It must be possible for NLP systems to find the relationships between a given word sense and other lexical entries. IITLEX includes both semantic relationships (e.g., a *kitten* is a kind of *cat*, specifically, a young cat) and syntactic relationships (e.g., the subject of the verb *dream* should be human or animate).

It is also essential to make the lexicon extensible. Lexical needs are always growing. We hope to add more information to IITLEX ourselves, from corpora and from the CED itself. Much of the information in the definition texts remains to be extracted and encoded in an explicit form.

## References

- [1] Ahlsweide, T.E., M. Evens, "Generating a Relational Lexicon from a Machine-

Readable Dictionary”, *International Journal of Lexicography*, 1 (3), 1988a, 214-237.

- [2] Ahlswede, T.E., M. Evens, “Parsing vs. Text Processing in the Analysis of Dictionary Definitions”, *Proceedings of 25th Annual ACL*, Buffalo, NY, 1988b, 217-224.
- [3] Ahlswede, T.E., D. Lorand, “Word Sense Disambiguation by Human Subjects: Computational and Psycholinguistic Applications”, *Proceedings of the ACL Siglex Workshop: Acquisition of Lexical Knowledge from Text*, Columbus, OH, 1993, 1-9.
- [4] Ahlswede, T.E., J. Anderson, M. Evens, S.M. Li, J. Neises, S. Pin-Ngern, J. Markowitz, “Automatic Construction of a Phrasal Thesaurus for an Information Retrieval System”, *Proceedings of RIAO88 (Recherche d'Information Assistée par Ordinateur)*, Cambridge, MA, 1988, 597- 608.
- [5] Alexander, D., W.J. Kunz, “Some Classes of Verbs in English”, Linguistics Research Project, Indiana University, Bloomington, IN, 1964.
- [6] Apresyan, Y.D., I.A. Mel’čuk, A.K. Žolkovsky, “Semantics and Lexicography: Towards a New Type of Unilingual Dictionary”, in F. Kiefer (ed.), *Studies in Syntax and Semantics*, Reidel, Dordrecht, 1970, 1-33.
- [7] Atkins, B.T., “Building a Lexicon: The Contribution of Lexicography”, in M. Bates, R.M. Weischedel (eds.), *Challenges in Natural Language Processing*, Cambridge University Press, Cambridge, 1993, 37-75.
- [8] Boguraev, B.K., “Building a Lexicon: The Contribution of Computational Lexicography”, in M. Bates, R.M. Weischedel (eds.), *Challenges in Natural Language Processing*, Cambridge University Press, Cambridge, 1993, 99-134.
- [9] Boguraev, B., B. Levin, “Models for Lexical Knowledge Bases”, *Electronic Text Research: Proceedings of the Sixth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, University of Waterloo, Waterloo, Ontario, 1990, 65-78.
- [10] Bridgeman, L., D. Dillinger, C. Higgens, P.D. Seaman, F. Shank, “More Classes of Verbs in English”, Indiana University Linguistics Club, Bloomington, IN, 1965.
- [11] Byrd, R.J., “Discovering Relationships among Word Senses”, *Proceedings of the Fifth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Waterloo, Ontario, 1989, 67-79.
- [12] Byrd, R.J., N. Calzolari, M. Chodorow, J. Klavans, M. Neff, O. Rizk, “Tools and Methods for Computational Lexicology”, *Computational Linguistics*, 13 (3/4), 1987, 219-240.
- [13] Calzolari, N., “The Dictionary and the Thesaurus can be Combined”, in M. Evens (ed.), *Relational Models of the Lexicon*, Cambridge University Press, Cambridge, 1988, 75-96.

- [14] Carver, D.J. (ed.), *Collins English Dictionary*, Collins Publishers, London, 1979.
- [15] Conlon, S. Pin-Ngern, M. Evens, "Can Computers Handle Adverbs?", *COL-ING'92*, Nantes, 1992, 1192-1196.
- [16] Conlon, S. Pin-Ngern, M. Evens, "A Lexical Database for Nouns to Support Parsing, Generation, and Information Retrieval", in S. Hockey, N. Ide (eds.), *Research in Humanities Computing*, Oxford University Press, Oxford, 1994, 75-87.
- [17] Conlon, S.P.N., M. Evens, T.E. Ahlsweide, R. Strutz, "Developing a Large Lexical Database for Information Retrieval, Parsing, and Text Generation Systems", *Journal of Information Processing and Management*, 29 (4), 1993, 415-431.
- [18] Dardaine, J., "Case Frames for a Lexical Database", *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Society Conference*, Starved Rock, IL, 1992, 102-106.
- [19] Evens, M., S. Pin-Ngern, T.E. Ahlsweide, S.M. Li, J. Markowitz, "Acquiring Information from Informants for a Lexical Database", *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, MI, 1989.
- [20] Evens, M., J. Dardaine, Y.F. Huang, S.M. Li, J. Markowitz, F. Rinaldo, M. Rinaldo, and R. Strutz, "For the Lexicon That Has Everything", *Proceedings of the Siglex Workshop*, Berkeley, CA, 1991, 179-187.
- [21] Fillmore, C.J., "The Case for Case", in E. Bach, R. Harms (eds.), *Universals in Linguistic Theory*, Holt, Rinehart, & Winston, New York, 1968, 1-88.
- [22] Grimshaw, J., *Verb Arguments*, MIT Press, Cambridge, MA, 1990.
- [23] Grimshaw, J., R. Jackendoff, *Report to the NSF on Grant IST-81-20403*, Department of Linguistics, Brandeis University, Waltham, MA, 1985.
- [24] Haynes, S., "A Tale of Two Parsers or Another CEDy Story", unpublished paper, Computer Science Department, Illinois Institute of Technology, Chicago, IL, 1993.
- [25] Hirst, G., "Why Dictionaries Should List Case Structure", *Advances in Lexicology: Proceedings of the University of Waterloo Centre for the New Oxford English Dictionary*, Waterloo, Ontario, 1986, 147-163.
- [26] Hoang, H.N.L., R. Strutz, M. Evens, "Lexical Semantic-Relations for Proper Nouns", *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Society*, Chesterton, IN, 1993, 26-30.
- [27] Householder, F., D. Alexander, P.H. Matthews, "Adjectives before That-Clauses in English", Indiana Linguistics Club, Indiana University, Bloomington, IN, 1964.
- [28] Householder, F., W. Wolck, P.H. Matthews, J. Tone, J. Wilson, "Preliminary Classification of Adverbs in English", Indiana University Linguistics Club, Bloomington, IN, 1965.

- [29] Jacobson, S., *Adverbial Positions in English*, A.B. Studentbok, Stockholm, 1964.
- [30] Jacobson, S., *On the Use, Meaning, and Syntax of English Preverbal Adverbs*, Almqvist & Wilksell International, Stockholm, 1978.
- [31] Jensen, K., J.L. Binot, "Disambiguating Prepositional Phrase Attachments by Using Online Dictionary Definitions", *Computational Linguistics*, 13 (3/4), 1987, 251-260.
- [32] Justeson, J.S., S. Katz, "Co-occurrences of Antonymous Adjectives and Their Contexts", *Computational Linguistics*, 17 (1), 1991, 1-20.
- [33] Lee, W., M. Evens, "Generating Cohesive Text Using Lexical Functions", *Proceedings of the International Workshop on Meaning-Text Theory*, Springer-Verlag, Berlin, 1992.
- [34] Levi, J.N., "Where Do All Those Other Adjectives Come From", *Proceedings of the Chicago Linguistics Society*, Chicago, IL, 1973, 332-345.
- [35] Levin, B., "Approaches to Lexical Semantic Representation", in Walker, D.E., A. Zampolli, N. Calzolari (eds.), *Proceedings of the Lexicon Workshop*, Linguistics Summer Institute, Stanford, CA, 1987.
- [36] Levin, B., "Building a Lexicon: The Contribution of Linguistics", in M. Bates, R.M. Weischedel (eds.), *Challenges in Natural Language Processing*, Cambridge University Press, Cambridge, 1993, 76-98.
- [37] Markowitz, J., S. Pin-Ngern, M. Evens, J. Anderson, S.M. Li, "Generating Lexical Database Entries for Phrases", in D.L. Berg (ed.), *Information in Text: Proceedings of the Fourth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Waterloo, Ontario, 1988, 115-127.
- [38] McCawley, J., "What Linguists Might Contribute to Dictionary Making If They Could Get Their Act Together", in P. Bjarkman and V. Raskin (eds.), *The Real-World Linguist*, Ablex, Norwood, NJ, 1986, 3-18.
- [39] Miller, G., R. Beckwith, C. Fellbaum, D. Gross, K. Miller, "Wordnet", *International Journal of Lexicography*, 4 (1), 1991, 1-75.
- [40] Morris, J., G. Hirst, "Lexical Cohesion, the Thesaurus, and the Structure of Text", *Computational Linguistics*, 17 (2), 1991, 21-48.
- [41] Nutter, J.T., E.A. Fox, M. Evens, "Building a Lexicon from Machine-Readable Dictionaries for Improved Information Retrieval", *Literary and Linguistic Computing*, 5 (2), 1990, 129-138.
- [42] Pin-Ngern, S., "A Lexical Database for English to Support Information Retrieval, Parsing, and Text Generation", Unpublished Ph. D. Dissertation, Computer Science Department, Illinois Institute of Technology, Chicago, IL, 1990.

- [43] Quirk, R., S. Greenbaum, G. Leech, J. Svartvik, *A Comprehensive Grammar of the English Language*, Longman, Harlow, Essex, 1985.
- [44] Rapp, C., D. Garfield, M. Evens, “Design of an Emotion Profiler Using SNePS”, in D. Kumar (ed.), *Current Trends in SNePS – Semantic Network Processing Systems*, Springer Verlag, New York, 1990, 145-152.
- [45] Sinclair, J.M. (ed.), *The Collins COBUILD English Language Dictionary*, London: William Collins & Sons, Ltd., London, 1987a.
- [46] Sinclair, J.M. (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, William Collins & Sons, Ltd., London, 1987b.
- [47] Smith, R.N., “Conceptual Primitives in the English Lexicon”, *Papers in Linguistics*, 18 (1), 1985, 99-137.
- [48] Sowa, J., “Using a Lexicon of Canonical Graphs in a Semantic Interpreter”, in M. Evens (ed.), *Relational Models of the Lexicon*, Cambridge University Press, Cambridge, 1988, 113-137.
- [49] Véronis, J., N.M. Ide, “Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries”, *COLING'90*, Helsinki, Vol. 2, 1990, 388-394.
- [50] Walker, D.E., “Knowledge Resource Tools for Accessing Large Text Files”, in G. Johannessen (ed.), *Information in Data: Proceedings of the University of Waterloo Centre for the New Oxford English Dictionary*, Waterloo, Ontario, 1985, 11-24.
- [51] Walker, D.E., R.A. Amsler, “The Use of Machine-Readable Dictionaries in Sub-language Analysis”, in R. Grishman and R. Kittredge (eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986, 69-83.
- [52] Walker, D.E., A. Zampolli, N. Calzolari (eds.), *Proceedings of the Lexicon Workshop*, Linguistics Summer Institute, Stanford, CA, 1987.
- [53] Wang, Y.C., J. Vandendorpe, M. Evens, “Relational Thesauri in Information Retrieval”, *JASIS*, 36 (1), 1985, 15-27.

# Visions of the Digital Library: Views on Using Computational Linguistics and Semantic Nets in Information Retrieval\*

Judith L. Klavans  
Columbia University  
*e-mail:* [klavans@cs.columbia.edu](mailto:klavans@cs.columbia.edu)

## Abstract

Whether there is a role for natural language processing techniques in information science has always been in question. Walker et al. 1977 report that the impact of linguistics, then a rapidly growing and vibrant field revolving around the study of language as a formal system, had surprisingly little impact on the field of information science, a field also revolving around the understanding of documents consisting largely of language. However, optimism and promise of results arising from collaborative efforts were also reported in Walker et al. 1977. In light of this optimism, this paper presents proposals for incorporating a semantic net derived semi-automatically from various machine-readable dictionaries into several aspects of the information retrieval task, including text indexing in order to cluster related documents, query expansion, and the construction of a browser as part of the human interface.

## 1 Natural Language Processing in Information Retrieval

Both skepticism and optimism can be found in viewing the complex relationship between information science and computational linguistics. In fact, whether there is a role for natural language processing techniques in information science has always been in question. Walker et al. 1977 report that, surprisingly, the impact of linguistics, then a rapidly growing and vibrant field revolving around the study of language as a formal system, had little impact on the field of information science, a field also revolving around the understanding of documents consisting largely of text.

Earlier, in 1970, the Committee on Linguistics in Documentation of the International Federation for Documentation, commissioned the preparation of a survey on the linguistic components of document analysis, description, and retrieval. The results of this survey appear in Sparck Jones and Kay 1973, and served as the impetus for the workshop organized and reported on by Walker et al. 1977. Issues that were examined

---

\*Portions of this paper arose from the Digital Library effort at Columbia University. In particular, I thank Kathleen McKeown, research director, for fruitful discussion of some of the ideas presented here. All inconsistencies or potential errors are, however, my own.

in the survey and which appear in Sparck Jones and Kay 1973 include: the relevance of linguistic theory and its relationship to the structure of knowledge, including phonology, morphology, syntax, and semantics; and discussion of how linguistic theory relates to the organization, storage, and transmission of information. Most importantly, the question of how the new field of computational linguistics could contribute and be affected by information science was examined. Also considered were topics which are still actively on current research agendas, such as automatic morphological and syntactic analysis, machine translation, machine-assisted translation, automatic indexing and abstracting, and the automatic preparation of indexes and thesauri.

The 1976 Workshop, leading to Walker et al. 1977, was specifically organized to gather together specialists in relevant sub-fields to develop a research agenda with a comprehensive plan. Although recognizing the realities and differences in the foci of linguistics, computational linguistics, and information science, the vision put forth in this volume was decidedly optimistic.

This optimism has persisted among a community of researchers within computational linguistics and information science who have followed the view that natural language processing and knowledge-based techniques do play a role in various aspects of the information retrieval (IR) field (Fox 1992). Research dating from the 60's (Salton and Lesk 1965, Salton 1966) explored knowledge-based methods for IR. Other research has focussed on the use of morphological information (Frakes 1992), linguistically sensitive collocational information (Smadja 1990), dictionary and thesaural methods (Krovetz and Croft 1989, Evans 1991, Liddy et al. 1993), syntax-based template and pattern matching (Fagan 1987), semantic representations, semantic nets (Voorhees 1993) and statistical and syntactic based phrasal indexing (Croft et al. 1991).

The key drawback in using such techniques are two. First, knowledge based techniques tend to be time and space intensive, thus slowing down performance. Secondly, semantic techniques tend to increase the number of overall retrievals, which in turn may degrade precision. However, research such as Belkin et al. 1993 suggests that multiple query based approaches are potentially highly promising; thus, such directions should be pursued, with the semantic net approach being part of a larger system incorporating different query and indexing functions. However, what is still lacking is a precise understanding of which knowledge based techniques will provide the best results.

With the understanding that some refinements may be necessary, this paper proposes using a semantic net to both index and query documents. The rationale for using the semantic net (rather than syntactically derived information) is the belief that most queries request documents organized around meaning. The definition of meaning in this context is not the more formal notion of logical form. Instead, we refer to a more utilitarian notion of semantics, including surface and deep aspects of lexical semantics.

## 2 Semantic Nets in Information Retrieval: A Proposal

Semantic nets have long been useful in other domains of artificial intelligence. Their properties and complexities are well studied although their drawbacks are also well-known (see Woods 1975). Wide-coverage semantic nets have been constructed both manually, as well as automatically and semi-automatically. Perhaps the most widely used manually constructed and publicly available general semantic net is WordNet (Miller

1990a, Miller et al. 1990b). Other manually constructed nets have either been small, domain specific, or not publicly available. Libraries have developed cataloging systems (e.g., the Library of Congress system) which can be considered a type of semantic net, covering subject areas. Other nets have been constructed automatically, using machine readable dictionaries or large texts.

**Semantic Nets and Knowledge-Based Approaches to IR** Knowledge-based approaches to IR are promising (Fox 1992), but the key is acquiring large knowledge bases. Knowledge can be obtained in many ways, including manual coding and entry, parsing (either rough or detailed), elicitation from experts, and from machine-readable lexicons and thesauri. The approach outlined here uses the semantic network schemata, where objects are indicated by nodes, and where links indicate relationships. We build on earlier work (Klavans et al. 1992) using semantic networks with hierarchical structures and inheritance. The knowledge bases we will use rely on machine-readable dictionaries, thesauri, and on structured analysis of large corpora. We expect our research to blend with other manually constructed hierarchies, when available, for a particular domain.

**Semantic Nets from Machine-Readable Dictionaries and Thesauri** Machine readable dictionaries have proven useful in the construction of large semantic nets (Amsler 1983, Chodorow, Byrd and Heidorn 1985, Ahlsweide et al. 1988, Wilks et al. 1989, Klavans et al. 1990). Klavans et al. 1990 demonstrated that a linguistically driven analysis of definitions permits the structuring of sophisticated nets, with some semantic inference possible over a large vocabulary. Our work has shown how most previous research on constructing hierarchies from dictionaries suffers from three problems: (1) only a limited portion of the dictionary data are accurately covered (fewer than 35% of noun definitions in Chodorow et al. 1985 are correctly structured); (2) a semantic network requires complex relationships beyond the simple IS-A link from the dictionary, and (3) the problem of polysemy is not properly handled.

Klavans et al. 1990 and 1992 demonstrated that more sophisticated techniques are required for semantic text analysis. We developed a more complex analysis for dictionaries to capture the richness of the lexicon as expressed implicitly in the dictionary. For example, from Webster's Seventh New Collegiate Dictionary, earlier techniques did not sensibly process definitions such as "ballot" (W71\$1b) "a sheet of paper", and "fascia" (W70\$3) "a sheet of connective tissue". These two defined nouns are not exactly related via "sheet" in the same way as "car" and "bus" are related via "vehicle". Similarly, a "band" (W\$0) is "a group of persons, animals, or things", whereas a "bevy" is "a group of animals and especially quail together". It would be strange to directly link "band" and "bevy" with IS-A links to "group". Instead, since they are certainly conceptually related via the "group" concept, what is required is the addition of semantically meaningful labels on links such as GROUP\_OF, UNIT\_OF, PART\_OF, MEASURE\_OF, MEMBER\_OF, and so on. In addition, a specification of group type is required, i.e., "group of animals" must be distinguished from "group of persons."

Many of these concepts are explicitly and implicitly found in dictionary definitions, although sometimes identifying the relation requires some clever linguistic analysis prior to automatic extraction. For example, the GROUP\_OF concept is differently expressed in the definition of "brown" (W2\$0) as "any of a group of colors between red and

yellow in hue”, than in “band” or “bevy” given above. Having addressed the problem of dictionary coverage and complex relationships to capture dictionary definition meanings, the third problem of polysemy still remains.

**Polysemy** Polysemy can be defined as several meanings or concepts represented by the same surface lexical item. To illustrate, a typical example of polysemy is the lexical item “bank”, which in the noun sense can mean “bank of a river”, or “the place where money is deposited...”. This example, although accurate, is misleading since there are many more subtle types of polysemy or lexical ambiguity that can degrade any natural language processing system. In the previous paragraph, the word “band” is vague as to whether it applies to people, animals, or things, to say nothing of other senses, such as “wedding band” or “rubber band.” The problem of polysemy can be said to be the most serious and difficult problem faced by any NLP application.

Why disambiguate polysemous items? The answer is that the semantic representation will have far less random noise without polysemy. For example, the twenty most frequent English nouns have an average of 7.3 senses. Similarly, the twenty most frequent verbs have an average of 12.4 senses. Krovetz and Croft 1992 found that for documents in a collection, each term has 3.7 to 4.6 mean senses, and for each query, each word has 6.8 to 8.2 mean senses. Since queries tend to be constructed with more general language, they tend to be even more vague than a document. The goal is to determine the *intended sense* of a term in order to use it correctly. As an example, a system must accurately guess whether the word “vehicle” intended to refer to the sense “vehicle for financial investment” or “vehicle for traveling from place to place.”

Recent research in solving the polysemy problem has relied on several sources, including MRDs and on statistical extraction of phrasal elements. What has been discovered is that, with accurate phrasal identification, ambiguity is sharply reduced. For example, “a band of oreoles” is unlikely to be confused with “a band of electrons”. Smadja and McKeown 1991 demonstrated that the use of the statistical analysis tool Xtract permitted the analysis of the lexical characteristics of documents in the financial domain for unambiguous term identification.

**Synonymy** Opposite to the problem of polysemy is the problem of synonymy. Synonymy can be defined as several words referring to the same concept, for example “attorney” and “lawyer”. MRDs and linguistically sensitive statistical processing of texts using tools such as Xtract can be used to expand a query by finding all synonymous equivalents of query terms. The danger of both, however, is the overexpansion of term sets leading to loosely related insufficiently similar terms. In order to control overexpansion, the polysemy problem must be solved. Thus, synonymy and polysemy interact to contribute to accurate sense-labelled representation of document meaning, and to sense-distinguished query formation.

**Tools and Resources** Voorhees 1993 reports on the effect of using a portion of the IS-A hierarchy from WordNet to disambiguate queries and to index documents. Although her results were negative, she comments that this may in part be due to the notion of synonym set (synset) in WordNet. Furthermore, although the IS-A relation is by

far the predominant one, using just this relation gives an inaccurate picture of lexical relations over documents. The Longman Dictionary of Contemporary English (40,000 headwords), Webster's Seventh (70,000 headwords), and the COBUILD dictionary are available on-line; Longman and Cobuild are enhanced with semantic and syntactic markings. In addition, other resources would enhance such research, such as a thesaurus (Collins and Roget are on-line), as well as corpus-based semantic nets (Smadja and McKeown 1991, Hatzivassiloglou and McKeown 1993). Other manually constructed hierarchies, when available, for a particular domain (Molholt and Goldbogen 1990), and other domain independent hierarchies could also be used.

## 2.1 Implementation and Approach

The input consists of documents and texts in any given domain. The first step is to preprocess text in fairly standard ways, such as sentence and word delimitation. These steps all involve research issues, such as the handling of hyphenation, abbreviations, and case. Next, text must be morphologically normalized, which also involves research questions such as inflectional and derivational affixation. Text is then ready to be tagged for part of speech. This permits noun recognition, verb-object recognition, and other part-of-speech dependent collocations to be extracted. For example, significant noun occurrences and noun sequences can be identified using statistical techniques such as discussed in Smadja and McKeown 1991, who demonstrated that documents in the financial domain can be lexically analyzed for unambiguous term identification. Such nouns and other significant single and multi-word terms can then be indexed with pointers into the semantic net. Consider a short excerpt of text from the legal domain:

"The issue arose in the context of a copyright dispute. RSO Records Inc. (RSO) owns a number of copyrights in sound recordings by the well-known pop artists the Bee Gees. Polydor Ltd. (Polydor) is the exclusive licensee of RSO under the copyrights in the United Kingdom, and pursuant to its exclusive license manufactures and markets records and cassettes reproducing those recordings in the United Kingdom." [document #638762]

Nouns such as "copyright" and "license", and phrases such as "copyright dispute", among others, emerge as significant within the larger document. Other concepts (such as "recording") are also important but not discussed here. Such nouns can be indexed according to: (1) the source document, [doc#] (2) location within the document [doc.location], and (3) links in the net [node.number(s)]. For example, the noun "copyright" (from "copyrights"), existing in given specific locations in a text and in the net, can be indexed to look like this: copyright [#638762;84;node.num.1,node.num.2], where the node numbers link to entry points in the semantic net. In this way, meaningful words and phrases can be processed and indexed. This serves two purposes: one is to organize text according to semantic criteria; the second is to provide a way to link texts with the semantic nets, and therefore allow a wider and richer search. Conversely, if the user or system specifies, such techniques can be used to narrow a search according to a generality measure for word meaning, explained below.

A portion of the net including "copyright" and "license" is given in the following figure, built from data in Webster's Seventh Dictionary. The definitions for "copyright"

and “license” from Merriam Webster’s Seventh Dictionary are first shown. The net represents some of the links among “copyright”, “license” and related concepts.

**copyright**

the exclusive legal right to reproduce, publish, and sell the matter and form of a literary, musical, or artistic work -- copyright adj

**license or licence**

1a permission to act

1b freedom of action

2a a permission granted by competent authority to engage in a business, occupation, or activity otherwise unlawful

2b a document, plate, or tag, evidencing a license granted

3a freedom that allows or is used with irresponsibility

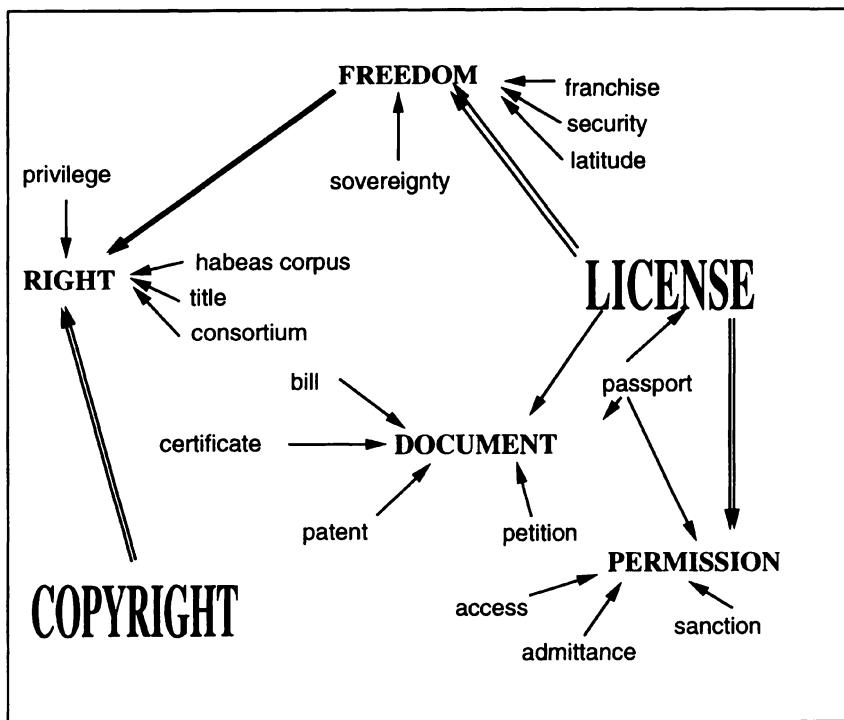


Figure 1: Simplified Semantic Net for “copyright” and “license”

Notice that “license” is a type of “permission” (W7:sense 1a,sense 2a) and a type of “freedom” (W7: sense 1b). A user will be able to see this relationship in the browser,

and pick other concepts related to “permission” and “freedom” for searching, such as “franchise” or “access”. A “license” is also a “document” (W7:sense 2b), and thus is related via the “document” concept to other words such as “passport”, “bill”, “patent”, “petition”, “registration”, and so on. In automatic searching, words higher in the net, those with more general meaning such as “action” or “thing”, will be weighted lower, since they tend to carry less specific content. This is the concept of a generality measure, derived from the net itself, based on frequency and net location. The impact of using such an approach over a key word search is that related concepts can be scanned for possible inclusion or exclusion from a query.

As demonstrated in Klavans et al. 1992, other complex relationships can be extracted from dictionaries, such as MEMBER\_OF, TYPE\_OF, GROUP\_OF, QUALITY\_OF, CONSISTS\_OF, and so on. As seen in the Figure, “copyright” and “license” are closely connected, via the “freedom” and “right” nodes. Thus, links can be drawn between highly abstract concepts, such as “freedom”, as well as concrete items, such as “document”. Within the structure of the net, the hyponyms of “document” include “patent”, “bill”, “certificate”, “petition”, and of course “license.”

Our method combines statistical and symbolic approaches to the problem of text analysis. The purely statistical approach functions by converting words into numerical objects and processing them as any statistical data. This method is useful for handling large volumes of text, and has been utilized successfully in most IR systems. However, it fails to capture the fact that the lexicon has properties peculiar to language, and particular to the nature of linguistic objects. The purely symbolic approach tends to require time-consuming manual input and, even more seriously, has exhibited limitations in scaling up to the level of efficiency and coverage required of broad-coverage IR systems. This research overcomes some of the limitations of the purely symbolic approach by using a large data set of lexical items, automatically extracted from large machine-readable dictionaries (MRDs) and corpora, and organized in a semantic network. Similarly, this approach overcomes some problems inherent in purely statistical methods by incorporating some of the uniquely linguistic properties of the lexical item. Statistical techniques and dictionary data can be used to deal with problems of word ambiguity (polysemy) and problems due to multiple words with the same or similar meaning (synonymy).

### 3 Conclusion

This paper has presented some ideas, inspired and motivated by earlier work on the role of linguistics in information retrieval as reported in Walker et al. 1977. The use of semantic nets may provide results in the indexing and querying of large texts, which will enable the development of more successful information retrieval systems which utilize linguistic and world knowledge. Since linguistics is the study of language, and since information retrieval depends on language, it is only natural that the two fields should interact, as observed in Walker et al. 1977.

## References

- [1] Ahlswede, T., J. Anderson, M. Evens, J. Neises, S. Pin-Ngern, J. Markowitz, “Automatic Construction of a Phrasal Thesaurus for an Information Retrieval System from a Machine Readable Dictionary”, *Proceedings of RIAO*, 1988, 597-608.
- [2] Amsler, R., “Machine-Readable Dictionaries”, *Annual Review of Information Science and Technology*, 1984, 19: 161-209.
- [3] Belkin N., “The Effect of Multiple Query Representations on Information Retrieval System Performance”, in R. Korfhage, E. Rasmussen and P. Willer (eds.), *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993, 338-346.
- [4] Chodorow, M., R. Byrd and G. Heidorn, “Extracting Semantic Hierarchies from a Large On-Line Dictionary”, in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, 1985, 299-304.
- [5] Croft, W.B., H.R. Turtle and D.D. Lewis. “The Use of Phrases and Structured Queries in Information Retrieval”, *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1991, 32-45.
- [6] Evans, D. E., K. Ginther-Webster, M. Hart, et al. “Automatic Indexing Using Selective NLP and First-Order Thesauri” *Proceedings of RIAO: Intelligent Text and Image Handling*, 1991, 624-643.
- [7] Fagan, J.L. “Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods.” PhD thesis, Department of Computer Science, Cornell University, 1987
- [8] Fox, E.A. Tutorial on Knowledge-Based Information Retrieval. Fifteenth International Conference on Research and Development in Information Retrieval, sponsored by the ACM SIGIR, 1992.
- [9] Frakes, W.B. “Stemming Algorithms”, in W. B. Frakes and R. Baeza-Yates (eds.), *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1992, 131-160.
- [10] Klavans, J.L., M.S. Chodorow and N. Wacholder. “From Dictionary to Knowledge Base via Taxonomy” *Proceedings of the Sixth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research*, University of Waterloo, Canada. 1990.
- [11] Klavans, J.L., M.S. Chodorow and N. Wacholder. “Building a Knowledge Base from Parsed Definitions”, in G. Heidorn, K. Jensen, and S. Richardson, (eds.) *Natural Language Processing: The PLNLP Approach*, Kluwer, New York. 1992.

- [12] Krovetz, R. and B. Croft. "Word Sense Disambiguation Using Machine-Readable Dictionaries", *Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1989, 127-136
- [13] Krovetz, R. and B. Croft. "Lexical Ambiguity and Information Retrieval", *Transactions of the ACL on Information Systems*, 1992, 10(2):115-141.
- [14] Liddy, E.D., W. Paik, and E.S. Yu. "Document Filtering Using Semantic Information from a Machine Readable Dictionary", *Proceedings of the ACL Workshop on Very Large Corpora*, 1993.
- [15] McKeown, K.M. and V. Hatzivassiloglou. "Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning", *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993.
- [16] Miller, G. Special Issue, "WordNet: An on-line lexical database", *International Journal of Lexicography*, 1990a, 3:4.
- [17] Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross and K.J. Miller. "Introduction to WordNet: An on-line lexical database", *Journal of Lexicography*, 1990b, 3(4): 235-244.
- [18] Molholt, P., G. Goldbogen. "The Use of Inter-Concept Relationships for the Enhancement of Semantic networks and Hierarchically Structured Vocabularies", *Proceedings of the Sixth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research, University of Waterloo*, 1990.
- [19] Salton, G. "Automatic Phrase Matching", in D.G. Hayes, (ed.), in *Readings in Automatic Language Processing*, Elsevier, New York. 1966, 169-188.
- [20] Salton, G. and M.E. Lesk "The SMART Automatic Document Retrieval System—an Illustration", *Communications of the ACM*, 1965, 8(6):391-398.
- [21] Smadja, F. "Retrieving Collocational Knowledge from Textual Corpora and Applications: Language Generation", PhD thesis, Computer Science Department, Columbia University, 1991.
- [22] Smadja, F. and K.M. McKeown, "Automatically Extracting and Representing Collocations for Language Generation", *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990, 252-259.
- [23] Sparck Jones K. and M. Kay, *Linguistics and Information Science*, Academic Press: New York, 1973.
- [24] Voorhees, E. "Using Wordnet to Disambiguate Word Senses for Text Retrieval", in R. Korfhage, E. Rasmussen and P. Willer (eds.), *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993, 171-180.

- [25] Walker, D. , H. Karlsgren, M. Kay, (eds.) *Natural Language in Information Science: Perspectives and Directions for Research*, Skriptor, Stockholm, Sweden, 1977.
- [26] Wilks, Y., D. Fass, C.-M. Guo, J. McDonald, T. Plate, and B. Slator. “A Tractable Machine Dictionary as a Resource for Computational Semantics” in B. Boguraev and T. Briscoe (eds.), *Computational Lexicography for Natural Language Processing*, Longman, London, 1989.
- [27] Woods, W. A. “What’s in a Link: Foundations for Semantic Networks,” in D.G. Bobrow and A. Collins (eds.), *Representation and Understanding*, Academic Press, New York, 1975, 35-82.

# Anatomy of a Verb Entry: from Linguistic Theory to Lexicographic Practice\*

Beryl T. Atkins  
Oxford University Press

Judy Kegl  
Rutgers University

Beth Levin  
Northwestern University  
*e-mail:* *b-levin@nwu.edu*

## Abstract

In the context of five learners' dictionaries, we examine the treatment of certain systematic relationships in the semantics and syntax of the English verb and find a lack of systematicity in this aspect of the lexicography. We present detailed evidence to support our criticism, including citations from a corpus of general English. We suggest that the discrepancies arise because of an inadequate representation of the native speaker's knowledge of the English verb system. Drawing on linguistic research into the verbal classification scheme of English, we construct a 'dictionary-neutral' summary of the semantic-syntactic relationships pertinent to the verb *bake*, designed to enable the lexicographer to make principled decisions about the content and presentation for a specific dictionary. On the basis of this summary we offer a revision of an entry for *bake*, showing that a theoretically motivated analysis, if clearly presented, will ease the lexicographer's task and improve the quality of the lexicography.

One striking property of lexical competence is how much a native speaker of a language knows about the syntactic potential of a word on the basis of little more than a knowledge of its meaning. This is pointed out by Hale and Keyser (1987), who consider the English verb *gally*, a whaling term, used in *the sailors galled the whales*. One native speaker of English, unfamiliar with this archaic verb, might assume from a wider context that *gally* means *see* (*the sailors saw the whales*), while another might take *gally* to mean *frighten* (*the sailors frightened the whales*). What is striking is that on the basis of these hypothesized meanings, the speakers will determine whether or not the verb *gally* is found in a variety of syntactic constructions. Specifically, Hale and Keyser consider

---

\*This paper is a slightly modified version of a paper with the same title that originally appeared in the *International Journal of Lexicography* 1:84–126 in the summer of 1988. It is reproduced here by kind permission of Oxford University Press.

the MIDDLE construction, involving a transitivity alternation where the subject of the intransitive middle use of a verb corresponds to the object of the transitive use. Thus the speaker who believes that *gally* means *see* would not allow the middle construction *whales gally easily* (cf. \**whales see easily*), while the speaker who interprets *gally* as *frighten* will find this construction perfectly acceptable (cf. *whales frighten easily*).<sup>1</sup>

Hale and Keyser argue that the middle construction is possible only with a semantically delimited class of verbs, those whose meaning involves a notion of change of state. They point out that *frighten*, *cut*, *split*, *open*, and *crush*, for example, have middles, but *see*, *consider* and *believe* do not. Thus the difference in the behavior of *gally* for the two speakers may be explained by their different interpretations of the sense of this verb: only the speaker who thinks that *gally* belongs to the class of verbs involving the notion 'change of state' (in this case, change in the psychological state of the experiencer of the emotion) will allow the middle construction.

This example shows that the meaning of a verb is used by the native speaker to predict a range of its syntactic properties. This process is possible because particular syntactic properties are tied to particular semantic classes of verbs. Once speakers use the meaning of a verb to establish its place in the organizational scheme of English verbs, they can predict its behavior. (The identification of the linguistically relevant aspects of meaning is the subject of much work in linguistics, as is the nature of the classification system of verbs.)

We shall now explore the implications of this view of lexical knowledge for lexicographic practice, drawing on the results of linguistic research into lexical organization, especially the work of the MIT Lexicon Project (Hale and Keyser 1986, 1987; Rappaport and Levin 1988; Rappaport, Levin, and Laughren 1988), which has paid particular attention to semantic-syntactic interdependencies of the type just described. In the first part of the paper, we identify inconsistencies in the way in which current learners' dictionaries deal with such interdependencies across a number of verbs which pattern together and suggest that the lack of any systematic approach in this area arises from an inadequate understanding of the semantic-syntactic properties of the verbs described. In the second part of the paper, we show how the results of research into lexical organization allow a clearer analysis of the properties of these verbs, and hence a more systematic treatment of them in the dictionary. Finally we attempt to apply this theory and suggest how one of the entries we examined may be improved without undue expansion or reorganization.

In our examination of the way dictionaries treat these particular linguistic relationships we have deliberately omitted dictionaries designed for native speakers, who are by definition assumed to know the lexical organization of their own language. In monolingual dictionaries the lexicographer need only present the sense(s) of the word in the way which best allows native speaker users to place that word as accurately as possible in the larger organizational schema of their language. It may justifiably be assumed that once this happens the users will be able to deduce the syntactic potential of the word, using the lexical knowledge which they as native speakers possess. This assumption is not valid for the users of dictionaries designed for non-native speakers, in whom the lexicographer cannot assume the same knowledge of the semantic-syntactic interdependencies of English. Such users require much more detailed and specific information in their dictionary, if they are to be able to use an English term as com-

---

<sup>1</sup>To save the curious a trip to the dictionary, *gally* means 'frighten, put to flight'.

petently and as flexibly as possible. The lexicographers whose task it is to assemble and present this material comprehensively and clearly are faced with making explicit their own native speaker lexical knowledge and intuitions. In our study of how they do this, we have chosen as representative works three monolingual and two bilingual dictionaries: *Oxford Advanced Learner's Dictionary* (OALD), *Longman Dictionary of Contemporary English* (LDOCE), *Collins COBUILD English Language Dictionary* (COBUILD), *Collins-Robert English-French Dictionary* (CREFD), and *Shogakukan Progressive English-Japanese Dictionary* (SPEJD).

## 1 Critical Assessment of Dictionary Entries

In this section, we shall explore the lexicographical fate of six types of semantic-syntactic interdependencies that are characteristic of the verb *bake*. To set the stage we examine three of them in some depth; subsequent interdependencies are discussed more briefly. The discussion of the first three interdependencies takes the following form: we first describe the phenomenon, next we critically assess how this phenomenon is handled in the entries for the verb *bake* across a variety of learners' dictionaries, and then we look at how one of these dictionaries handles the same phenomenon in other verbs. We want to note at the outset of our investigation that we have restricted ourselves to a subset of facts which we feel are central to any account of the verb *bake*, although we recognize that there are many other properties of this verb that are not covered in this paper.

Our study of the verb *bake* highlights the treatment in learners' dictionaries of the rich set of semantic-syntactic interdependencies associated with this verb. These interdependencies are highly familiar to native speakers of English but are particularly problematic to learners. Interdependencies are difficult to code in dictionaries and are often signalled by the careful juxtaposition of example sentences in conjunction with a number of implicit cues in the organization and presentation of sense distinctions. Frequently, interdependencies become apparent when uses of a verb that involve alternate expressions of the verb's arguments within a single sense are compared and contrasted. Often these alternations involve transitive and intransitive uses of a verb; we refer to such transitive/intransitive pairs as TRANSITIVITY ALTERNATIONS. Two pairs of transitivity alternations exhibited by the verb *bake* on the 'cooking food' sense appear in (1) and (2).<sup>2</sup>

- (1)    a.    'As a change from moules marinières, try baking them in their shells in hot ash.' (transitive)
- b.    'While the soufflé was baking and she was searching around for something to do . . .' (intransitive)
- (2)    a.    'Every morning they baked their own baguettes and croissants.' (transitive)

---

<sup>2</sup>Because this paper is aimed at both lexicographers and theoretical linguists, the examples in this paper may seem more true-to-life than those linguists typically encounter: they are. Where possible, all examples (except those which form part of existing dictionary entries) are drawn from the Birmingham Collection of English Text (BCET) Corpus. All corpus examples are placed in inverted commas, while composed examples, where used, have no quotes.

- b. ‘As we baked, we talked a great deal.’  
(intransitive)

Notice that (1) and (2) involve extremely different intransitives. In pair (1) the object of the transitive, like the subject of the intransitive, denotes the entity that is cooked, whereas in pair (2) the subjects of the transitive and intransitive both denote the person who carried out the cooking. Notice also that the meanings of the transitives in these two sets of examples differ subtly. We will argue that (1) involves *bake* in the sense where it means ‘to change the state of something by cooking it in dry heat’ and that (2) involves *bake* in the sense where it means ‘to create something by cooking in dry heat’. We also propose that these semantic differences correlate with differences in the syntactic realization of arguments when the verb participates in transitivity and other alternations.

## 1.1 The Causative/Inchoative Alternation

The examples in (1) illustrate a transitivity alternation, the CAUSATIVE/INCHOATIVE ALTERNATION. Both the transitive and intransitive sentences describe a particular change of state in some entity, typically food. We refer to the argument denoting the entity that undergoes the change of state as the PATIENT. The intransitive use of *bake* simply describes the fact that the patient undergoes a change of state, without specifying how the change of state comes about. This intransitive use of *bake* alternates with a transitive use which explicitly requires the expression of the AGENT who brings about the change of state.

### 1.1.1 Signalling the Causative/Inchoative Alternation

Let us begin by taking a look at how this transitivity alternation and the associated difference in meaning are signalled in the OALD entry for *bake*:<sup>3</sup>

**bake** ... *vt, vi* [VP6A,22,2A,C] 1 cook, be cooked, by dry heat in an oven: ~ *bread/cakes; ~d beans*. *The bread is baking/being ~d.* 2 make or become hard by heating: *The sun ~d the ground hard. Bricks and earthenware articles are ~d in kilns.*

This entry signals the fact that *bake* has transitive and intransitive uses. Transitivity is indicated in two ways: through the parts of speech (“*vt, vi*”) and the syntactic patterns (“[VP6A,22,2A,C]”). The entry also gives examples of both transitive and intransitive uses of the verb that reflect the causative/inchoative alternation. The two uses are included under a single sense category. Lumping these two uses together under a single sense category, despite their difference in transitivity, highlights the fact that they both involve the ‘cooking (food)’ sense of *bake* and distinguishes them from the transitive and intransitive uses connected with a second sense of *bake*, the ‘hardening (clay)’ sense, which we will not consider in this paper.

---

<sup>3</sup>Within the article, we often quote only the relevant part of entries. The full text of each dictionary entry we refer to can be found in Appendix II of the 1988 version of this article.

This entry virtually ignores the creation sense of *bake* seen in (2). Although it lumps together the transitive and intransitive ‘cooking’ uses which flag *bake* as a member of the class of change of state verbs, the entry never explicitly acknowledges that *bake* is a change of state verb. Instead, it conveys this information indirectly by showing that *bake* participates in the causative/inchoative alternation, a well-documented transitivity alternation known to be associated with the class of change of state verbs. A representative list of the members of this verb class appears below.

- (3) bend, blacken, boil, break, broil, brown, burn, close, cool, cook, darken, defrost, dye, drop, dry, evaporate, freeze, fry, gasify, grill, grow, harden, heat, melt, move, open, poach, redden, ripen, roast, roll, sharpen, shorten, soften, split, stabilize, steam, stew, stiffen, thicken, thin, toast, whiten, ...

*Bake*’s participation in the causative/inchoative alternation is signalled in several ways in OALD. The first cue is the part of speech label and grouping of syntactic codes, which suggest that *bake* participates in some type of transitivity alternation. Second, the collapsing of two definitions into a single sense category (“cook, be cooked”) links the two uses. Still, the learner must be informed more explicitly of the relations between the causative and inchoative uses, namely that the subject of the intransitive use of *bake* bears the same semantic relation to the verb as does the object of the transitive use. The syntactic codes simply indicate transitivity and do not indicate whether the subject of the intransitive bears any relation to the subject or the object of the transitive. This task is left to the examples, where this fact is shown by the careful use of parallel noun phrases (e.g., *bread*) to denote the patient argument. The two main parts of the definition “cook” and “be cooked” are also intended to transmit this information to the users, who require considerable skill if they are to match “be cooked” with the intransitive use and “cook” with the transitive use. The learner also needs information about how these two uses differ. The transitive use involves a second participant, the agent bringing about the action, which is not expressed in the intransitive use.

### 1.1.2 The Causative/Inchoative Alternation Across Entries for *Bake*

The LDOCE (1987) entry for *bake* parallels the OALD entry in several respects.

**bake** ... v [I;T] to (cause to) cook using dry heat in an OVEN: *to bake bread|The bread is baking.|baked potatoes in their jackets (=with the skin on)—see COOK (USAGE).*

The entry again lumps together the transitive and intransitive uses of *bake* in the ‘cooking’ sense under a single sense category. Furthermore, it indicates that the verb has transitive and intransitive uses in this sense through the syntactic codes ([I;T]) and gives an example of each, again employing parallel noun phrases to highlight the relationship between them. The wording of the LDOCE definition is also intended to bring out the common core of meaning shared by the two uses. Where the OALD definition is disjunctive, using the two phrases “cook” and “be cooked,” which both share the same modifier “by dry heat in an oven,” LDOCE makes use of a special device to signal what the two uses have in common and where they differ. Part of the definition, the

string “cause to,” is enclosed in parentheses to indicate that it is optionally a part of the definition. Thus the definition encapsulates the two definitions below:

- (4) a. to cook using dry heat in an OVEN  
b. to cause to cook using dry heat in an OVEN

By collapsing the two definitions into one, this device captures the similarities and differences between the transitive and intransitive uses succinctly and perspicuously. But LDOCE leaves it to the user to determine the correct reading of this definition: the presence of “cause to” signals the definition of *bake* associated with the transitive use, while the variant without this phrase is presumably the definition of the intransitive use. In this instance, this device may simply prove confusing to the learner. *Cook* shares with *bake* the property of participating in the causative/inchoative alternation, itself having transitive and intransitive uses related by the causative/inchoative alternation (*Jess cooked the potatoes./The potatoes were cooking.*). This property means that the learner must recognize that it is the intransitive use of *cook* that is being referred to in the LDOCE definition.<sup>4</sup>

The COBUILD dictionary again treats the two cooking uses of *bake* as belonging under one sense category.

**bake** . . . When you **bake** food or when it **bakes** it is cooked in an oven without using extra liquid or fat. EG *She said she would bake a cake to celebrate. . . I cleaned the kitchen while the bread was baking.* ◊ **baked**.  
EG . . . *baked potatoes*.

COBUILD indicates a variety of additional information about each entry, including syntactic information, in a separate ‘marginal’ column. The marginal material associated with the entry for *bake* includes the syntactic code “V-ERG”; this is a special syntactic code used in COBUILD to indicate that a verb has transitive and intransitive uses ‘linked’ in the specific semantic relationship that characterizes the causative/inchoative alternation. The examples in the COBUILD entry illustrate transitive and intransitive uses, but they do not use parallel noun phrases as subject and object, reflecting the fact that such examples were not to be found in the corpus from which the examples are drawn. But the definition itself, which uses full sentences, makes clear that the object of transitive *bake* and the subject of intransitive *bake* bear the same semantic relation to the verb by indicating that “food” qualifies as the subject of the intransitive and the object of the transitive. Despite the constraints that the use of a corpus imposes on the choice of examples, note that the examples have been chosen so that the object of the transitive example and the subject of the intransitive example are nouns that are prototypical patients of *bake*.

Having examined three monolingual dictionaries for learners, let us now consider how two bilingual dictionaries, which by definition are also for learners, deal with the same facts. We begin with the CREFD.

**bake** . . . 1 vt (a) (*Culin*) faire cuire au four. **she ~s her own bread** elle fait son pain elle-même; **to ~ a cake** faire (cuire) un gâteau; **~d apples/potatoes**

<sup>4</sup>Interestingly, the *Collins English Dictionary* even goes as far as exploiting the ambiguity of *cook* in writing its entry for *bake*: “to cook by dry heat in or as if in an oven.”

pommes *fpl*/pommes de terre au four; ~d Alaska omelette norvégienne; ~d beans haricots blancs à la sauce tomate; V half.

(b) *pottery, bricks . . .*

vi (a) [*bread, cakes*] cuire (au four).

(b) she ~s every Tuesday (*makes bread*) elle fait du pain le mardi; (*bakes cakes*) elle fait de la pâtisserie tous les mardis.

The organization of the entry in this dictionary is quite different from the first three. The policy here is to split transitive and intransitive uses of a verb into separate sections ('grammatical categories' in the terms of the compilers of this dictionary), making it even more difficult to show explicitly any semantic-syntactic relationships because the transitive use of *bake* is under sense ('semantic category') 1(a), while the intransitive use is under sense 2(a). Again the lexicographer has signalled the relation between the two uses by using the same nouns (*cake* and *bread*) as instances of the prototypical object in the transitive example and as instances of the prototypical intransitive subject in the metalanguage of the intransitive definition.<sup>5</sup>

Finally, we turn to a second bilingual dictionary, the SPEJD. This dictionary, like the other bilingual, the CREFD, again structures its entry for *bake* according to transitivity. The uses we are concerned with are found under sense category 1 of both the transitive and intransitive parts of the entry (but unlike the CREFD, sense category 1 of the intransitive applies both to the 'cooking' and 'hardening' senses of *bake*, with splits inside the sense category to distinguish among the two). Like the other entries we have examined, this entry includes examples of both the transitive and intransitive uses. The noun *cake* appears as both the subject of the intransitive example ('The cake is still baking') and the object of the transitive example ('~ bread [cake, meat] in an oven'), serving as a cue to the user that there is a relation between them. This dictionary also explicitly indicates in Japanese what kind of things each sense of *bake* pertains to, giving a list of the relevant items within angle brackets. Thus under transitive sense category 1, the phrase in angle brackets indicates that this use applies to things such as bread. Within intransitive sense category 1, each of the separate subsenses (delimited by semi-colons) is preceded by an indication of what it applies to. This is necessary because the 'cooking' sense of *bake* does not have the same Japanese translation as the 'hardening' sense, but both are lumped under the same sense category.

To summarize, we have looked at five major learners' dictionaries and have seen that in each case the lexicographer is aware (i) that the causative/inchoative alternation exists, and (ii) that the characteristic properties of this alternation form a necessary part of a comprehensive description of the verb *bake*. As shown by the choice of words in the definition, the language of the examples, and the metalinguistic material, the lexicographers clearly believe that the systematic relation between the uses must be conveyed to the user since they are an essential component of knowing this verb.

The entries for *bake* in all of the dictionaries examined reflect the lexicographers' awareness of the causative/inchoative alternation, although this insight is not adequately conveyed to the learner, who will undoubtedly not have the skills required to extract a unified generalization from a set of disparate cues.

---

<sup>5</sup>We are not concerned with the treatments of the French counterparts of *bake* on the French/English side of the CREFD, because we are interested only in the way the English facts are presented to the dictionary user.

### 1.1.3 The Causative/Inchoative Alternation in the Context of the Entire Dictionary

Verbs that share common components of meaning usually share common syntactic properties. Such verbs tend to participate in the same transitivity alternations; that is, they have transitive and intransitive uses that are related in the same way. For instance, the causative/inchoative alternation exhibited by *bake* exists as part of a wider pattern in English; it is seen with other verbs of change of state, and, specifically, other verbs of change-of-state-by-cooking, among them *boil*, *broil*, *brown*, *fry*, *grill*, *poach*, *roast*, *steam*, *stew*, *toast*, and, as we have seen, *cook* itself. We focus on four additional verbs from this list, whose participation in the causative/inchoative alternation is illustrated with the examples below from the BCET Corpus.

- (5)    a.    'He may be husky but he can't boil an egg.' (transitive)  
      b.    'I led him to the sink while his two eggs boiled.' (intransitive)
- (6)    a.    'A second man was frying six eggs in a bent frying pan.' (transitive)  
      b.    'They had fried to a crisp.' (intransitive)
- (7)    a.    'In honour of this signal occasion Mrs. Green roasted a turkey stuffed with rice and garlic.' (transitive)  
      b.    'while the bird roasted he built a rough bush shelter.' (intransitive)
- (8)    a.    'You can use fresh or frozen fruit that you have stewed for the rest of the family.' (transitive)  
      b.    'All these elements stewing in a huge cauldron were bound to boil over.' (intransitive)

Ideally, the lexicography throughout the dictionary should bring out the fact that the behavior of these verbs reflects a larger pattern. We now return to OALD to see whether the lexicographers have recognized this pattern in the entries of cooking verbs, and, if so, how it is handled.

It appears that the lexicographers realized that there is a pattern here. The phrase "cook, be cooked," used in *bake*'s definition to signal its participation in the causative/inchoative alternation, is repeated in the definitions of other cooking verbs. This strategy is extended, for example, to the OALD entries for *fry*, *roast*, and *stew*.

**fry<sup>1</sup>** . . . *vt, vi* . . . [VP6A,2A] cook, be cooked, in boiling fat; *fried chicken*.  
*The sausages are frying/being fried.*

**roast** . . . *vt, vi* [VP6A,2A] (of meat, potatoes, etc) cook, be cooked, in a hot oven, or over or in front of a hot fire, e.g. on a turnspit, the meat, etc being basted periodically with the fat and juices that come out; ~ a joint. *The meat was ~ing in the oven. You've made a fire fit to ~ an ox*, a very large, hot fire.

**stew<sup>1</sup>** . . . *vt, vi* [VP6A,15A,2A,C] cook, be cooked in water or juice, slowly in a closed dish, pan, etc; ~ed chicken/fruit; ~ing pears, suitable for ~ing but not for eating uncooked.

The definitions of these verbs differ only in containing the appropriate description of the manner of cooking.

We would expect that, on analogy with *bake*, *fry*, *roast* and *stew*, the OALD entry for *boil* would also include the phrase “cook, be cooked” with the appropriate manner description. But this is not the case:

**boil<sup>2</sup>** . . . *vi, vt* **1** [VP2A,B,C,D] (of water or other liquid, also of the vessel that contains it and of the contents of the vessel) reach the temperature at which change to gas occurs; bubble up: *When water ~s it changes into steam. The kettle is ~ing. The potatoes are ~ing. Don't let the kettle ~ dry. Let the vegetables ~ gently. . .*

**3** [VP6A,22] cause water or other liquid to ~; cook in ~ing water: *We ~ eggs, fish and vegetables. Please my egg for three minutes. I like my eggs ~ed hard. My brother prefers soft-~ed eggs.*

The transitive and intransitive uses of *boil* are split into two different sense categories. “*The potatoes are boiling*” is found in sense category **1**, while “*please boil my egg for three minutes*” is in sense category **3**. Despite the various types of possible subjects for the intransitive use of this verb (“*kettle*”, “*water*”, “*potatoes*”), there seems to be no good reason for treating the two variants of this alternation as two separate sense categories. It might be objected that collapsing the transitive and intransitive uses would result in too clumsy an entry; we would respond that it would have been more helpful to split according to the type of subject/object rather than according to transitivity. Such a split would yield three sense categories: (i) (of water) to reach or cause to reach the temperature at which change to gas occurs, (ii) (of the vessel) to reach or cause to reach the temperature at which change to gas occurs, and (iii) (of food or solids cooked in liquid) to cook or be cooked in liquid which has reached the temperature at which change to gas occurs. (Incidentally, the entries in LDOCE and COBUILD, also monolingual learners’ dictionaries, do in fact treat *boil* along the lines just suggested.)

We have just seen that the same phenomenon, the causative/inchoative alternation, has been handled in different ways within a single dictionary. This inconsistent treatment hinders the learner’s ability to recognize that certain semantically related groups of words share certain syntactic properties. Our overview of the five learners’ dictionaries shows that the causative/inchoative alternation is not consistently and comprehensively treated in any of them, even though it is clear that the lexicographers were aware of its existence and of its importance.

## 1.2 The Indefinite Object Alternation

It is particularly important that the causative/inchoative alternation should be handled well, because *bake* shows more than one intransitive use. The learner needs to be made aware of the difference between the two distinct types of intransitives which are associated with two distinct transitivity alternations: the causative/inchoative alternation and the indefinite object alternation. In the first alternation, which we have just examined, the object of the transitive verb has the same semantic relation to the verb as the subject of the intransitive. The second, which we come to now, is the INDEFINITE OBJECT ALTERNATION, exemplified below with corpus examples (repeated from (2)):

- (9) a. ‘Every morning they baked their own baguettes and croissants.’ (transitive)  
       b. ‘As we baked, we talked a great deal.’  
             (intransitive)

In this alternation, the subjects of both the transitive and intransitive uses of the verb express the agent of the action. In the intransitive indefinite object use no direct object is expressed, but the sentence still implies that some foodstuff was cooked.<sup>6</sup> The exact nature of what is cooked is left unexpressed and unspecified, although it is generally understood to be a flour-based product such as bread, cake, or pastry. Thus the understood object of this use of *bake* has a narrower interpretation than the range of possible objects of the transitive use.

The uses of *bake* found in both the causative/inchoative and indefinite object alternations share a common core of meaning: ‘cooking in dry heat’. But while the causative/inchoative alternation is associated with one sense of *bake*, the change of state sense, ‘to change the state of something by cooking it in dry heat’, the indefinite object alternation is associated with a second sense of *bake*, ‘to create something by cooking in dry heat’. Just as we demonstrated that English possesses an entire class of change of state verbs, we can also show that English possesses a class of creation verbs, including:

- (10) build, carve, cast, cook, crochet, embroider, grow, knit, sculpt, sew, spin, weave, whittle, . . .

Unlike the causative/inchoative alternation, the indefinite object alternation is not found with verbs of change of state, but it is characteristic of a number of classes of verbs, among them verbs of creation, verbs of ingesting (*eat, drink, sip, . . .*), and verbs describing various occupation-related activities (*crochet, plow, weed, sweep, iron, hunt, type, write, file, . . .*). The change of state sense of *bake* found in the causative/inchoative alternation simply describes a change in the state of some entity brought about by cooking in dry heat. The creation sense of *bake* found in the indefinite object alternation is more complex: it describes the creation of an object by means of change of state *bake*.

### 1.2.1 The Indefinite Object Alternation across Entries for *Bake*

The CREFD recognizes that *bake* is found in the indefinite object alternation, as is apparent from the selection of examples in the part of the entry cited above: “*she bakes her own bread*” (in category 1(a)) and “*she bakes every Tuesday*” (in category 2(b)). The two intransitive uses of *bake* are differentiated by being placed in different categories: the indefinite object (intransitive) variant is in category 2(b) and the intransitive variant of the causative/inchoative alternation is in category 2(a).

There is no reference to the fact that *bake* participates in the indefinite object alternation in OALD, LDOCE, or SPEJD. The COBUILD lexicographer indicates an awareness of this possibility by the marginal material associated with the *bake* entry, where the codes “v” as well as “V-ERG” are noted. However, only the “V-ERG” uses are exemplified in the entry, and the reader is left with the almost impossible task of deducing that the

---

<sup>6</sup>In this paper we use the terms TRANSITIVE and INTRANSITIVE in a purely syntactic sense. Thus we consider the use of *bake* in (9b) to be intransitive despite the fact that this sentence has an understood object.

“v” refers to the indefinite object use. This use of the verb *bake*, which is not discussed in four of the five dictionaries we are looking at, is not an overly obscure usage, as is proved by the presence of six instances in the 290-odd citations of the verb *bake* in the BCET Corpus, example (9b) above and the following examples:<sup>7</sup>

- (11) a. ‘Thank God I baked this morning.’
- b. ‘Mrs. Burns was baking: it was Lionel’s day off.’
- c. ‘She arrayed the house, cooked and baked.’
- d. ‘We baked in a reflector oven—isn’t that terrific?’
- e. ‘... washed and cooked and baked and shopped ...’

### 1.2.2 The Indefinite Object Alternation in the Context of the Entire Dictionary

As mentioned above, *bake* is not the only English verb to participate in the indefinite object alternation. Various creation verbs, including *embroider* and *sew*, as well as other cooking verbs, show this property:

- (12) a. ‘Mrs. Babcock is embroidering a sampler.’
- b. ‘I didn’t even know she could embroider.’
- (13) a. ‘She was sewing her own dresses.’
- b. ‘She made herself useful. She sewed and mended.’

Given the failure of so many dictionaries to note the alternation in *bake*’s entry, let us look outside the class of cooking verbs in one dictionary, the CREFD, to see how the alternation is treated. This dictionary does not signal this alternation consistently. It recognizes only the transitive use of *embroider*, though it includes both transitive and intransitive uses of *sew*. Looking through the other four dictionaries we find similar inconsistencies in their treatment of these verbs and other verbs which function in the same way.

We have seen that the indefinite object alternation, while hinted at in the entries for *bake*, is not signalled consistently within a specific dictionary, nor is it fully described in any of the dictionaries under consideration.

### 1.3 The Benefactive Alternation

We have already exhausted the properties of *bake* that are recorded in the learners’ dictionaries we are studying. Are there others that are not treated at all in any of these dictionaries? There are. One is the BENEFACTIVE ALTERNATION, exemplified from the BCET Corpus in (14).

- (14) a. ‘Jennifer had baked a special cake for Alexander.’ (transitive + *for*)
- b. ‘Bake us plenty of bread, therefore, tonight.’  
(ditransitive)

---

<sup>7</sup>See the appendix for a detailed breakdown of the occurrences of *bake* in the BCET Corpus.

Notice that these examples involve alternate expressions of the arguments of *bake* but the alternation here is not a transitivity alternation; neither variant is intransitive.

Unlike the uses of *bake* examined so far, the uses in (14) involve a third participant besides the two we have already encountered (agent and patient). This third participant is the BENEFICIARY of the action, *Alexander* in (14a) and *us* in (14b). What distinguishes the two variants in (14) is the different mode of expression of the benefactive and patient noun phrases. In (14a), the benefactive noun phrase is expressed in a prepositional phrase headed by *for*, while in (14b) it is expressed as the first object of a double object construction. Concomitant with this, the verb is transitive in (14a) and ditransitive in (14b) and the patient argument is expressed as the object of the verb in (14a) and as the second object in (14b).

The benefactive alternation shown by *bake* must be distinguished from the DATIVE ALTERNATION shown by verbs such as *give*, which also involves an alternation between a transitive use of a verb plus prepositional phrase and a ditransitive use. Contrast the dative alternation in (15) with the benefactive alternation in (16).

- (15) a. Jennifer gave a present to Alexander.  
b. Jennifer gave Alexander a present.
- (16) a. Jennifer baked a cake for Alexander.  
b. Jennifer baked Alexander a cake.

The two alternations are differentiated by the interpretation of the first object in the double object construction. It is the recipient of the action in the dative alternation and the beneficiary of the action in the benefactive alternation. This difference is reflected in the prepositions associated with the prepositional variant of each alternation: *to* in the dative alternation and *for* in the benefactive alternation.

We claim that the benefactive alternation is found with verbs belonging to such classes as creation and obtaining (*make me a dress*, *buy me a book*), but not with verbs of pure change of state (\**break me a glass*). We have noted that verbs of cooking such as *bake*, *boil*, and *fry* are fully-fledged members of both the class of change of state verbs and the class of creation verbs. We believe that the benefactive alternation is associated only with the creation sense of these verbs. No one would doubt that an example like *bake me a cake* involves the creation sense of *bake*. Clearly the cake in this sentence is created from various raw materials such as eggs, butter, sugar, and flour and does not exist as an entity prior to the baking process. However, the existence of examples like *boil me an egg* or *fry me an egg* seems at first glance problematic, precisely because the egg in each of these examples exists as an entity before it is boiled or fried. Despite this, we feel that the entity which exists prior to the boiling and frying is distinct from the edible entity ‘created’ by these processes. Consider for example, *fry me a sunny-side up egg*, where the raw material is a raw egg and the result is a very specific type of cooked egg. As the benefactive alternation is a particularly complex property of *bake*, it is a particularly useful one to encode in a learners’ dictionary entry.

### 1.3.1 Signalling the Benefactive Alternation

None of the five dictionaries we have examined notes that *bake* participates in the benefactive alternation. This fact is particularly striking when we remember that this

alternation is explicitly encoded in many dictionaries, because it is found with such a wide range of verbs. Take for example the SPEJD, which has devised a schematic method of indicating just such a fact to the user. Sense category 2 of the transitive *boil* entry in SPEJD reveals that this dictionary is equipped with two formal ways of informing the user of the existence of the benefactive alternation: schematically as in (17) and by means of parallel examples linked by “=” as in (18).

- (17) [boil A B / boil B for A]

- (18) She boiled *me* some tea [=She boiled some tea *for me*]

Both devices pair the double object variant with a *for* prepositional variant in order to signal to the dictionary user that the double object construction receives a benefactive and not a dative interpretation.<sup>8</sup> The same device is used to indicate that *cook* shares this property. However, although the mechanism for doing so clearly exists, this dictionary fails to note that *bake* also can undergo this alternation, nor that this is possible for *fry* or *stew*.

LDOCE (1978) also has a formal method of indicating, by means of the code “D1”, when a verb is found in the double object construction. Indeed the LDOCE entry for *boil* sense category 2, which contains the example “*Shall I boil you an egg?*,” is explicitly coded as “[D1 (for);T1]”, showing by the inclusion of *for* in brackets that this example could be paraphrased as *Shall I boil an egg for you?* and not as the ungrammatical \**Shall I boil an egg to you?*. However, although the dictionary designers have foreseen the need to indicate this alternation, the lexicographers do not indicate it systematically through the dictionary, witness the entries for *bake*, *cook*, *fry* or *stew*. Indeed in the second edition of LDOCE (1987), this formal method of indicating the benefactive alternation has been abandoned.

OALD’s description of Verb Pattern 12B (associated with ditransitive verbs) includes the words: “In this pattern the indirect object is equivalent to a prepositional adjunct with ‘for’, as in [VP13B]” (p. xxiii). One of the examples for [VP12B] is “*She made herself a new dress*,” and for [VP13B] “*She made a new dress for her youngest daughter*. ” Thus we see that OALD, like LDOCE and SPEJD, has a formal way of indicating that a verb has the property of participating in the benefactive alternation. This fact is signalled for the verb *cook*, where the pattern is noted although not exemplified. However, OALD does not deal with this facet of verb behavior systematically. There is no mention of it in the entries for *bake*, *fry*, *boil*, *stew*, or other similar verbs.

COBUILD, in the grammatical notes at “V+O+O” and “V+O+A”, states specifically that ditransitive verbs exist and that some of these verbs have the possibility of a transitive variant with a prepositional phrase headed by *for*. This coding is used in the dictionary, but not systematically or exhaustively. For example, *cook* is described in the marginal material as “V-ERG or V: also V+O+O, V+O+A(FOR)”. However, no mention is made of this fact at *bake*, *fry*, *boil*, or *stew*.

To end our survey, CREFD has no means of indicating systematically and formally the existence of ditransitive verbs and their prepositional variants. In some entries, this

<sup>8</sup> SPEJD actually has a third formal way of indicating the alternation, the use of the verb class codes ‘III’ and ‘IV’, which encode the syntactic frames associated with this alternation (V NP PP and V NP NP). But this device does not provide information about the preposition involved in the V NP PP frame, a prerequisite for realizing that *bake* participates in the benefactive alternation.

fact is noted by means of metalinguistic notes (e.g., “for sb”) or explicitly stated in the examples, but this is not the case for *bake*, *fry*, *boil*, *stew*, *cook*, and other verbs that share these properties.

To conclude our discussion of the benefactive alternation, we note yet again that some dictionaries have a formal means of showing that a verb participates in this alternation and that others do not. However, even those which have the means at their disposal for doing this systematically do not indicate the alternation wherever it occurs.

In identifying three further alternations in the context of the verb *bake*, it would serve no purpose to expound at length all the relevant facts in the amount of detail which we have included in Sections 1.1, 1.2, and 1.3 above. Therefore, we shall briefly outline and exemplify the remaining alternations, and simply note that none of them is treated systematically in any of our five representative dictionaries.

## 1.4 The Material/Product Alternation

The MATERIAL/PRODUCT ALTERNATION, like the benefactive alternation, involves the alternate expression of certain arguments of the verb *bake* without a change in transitivity.

- (19) a. ‘They baked unleavened bread from the dough.’  
b. ‘the new corn, whether raw, roasted, or baked into bread . . .’

This alternation can be seen more clearly when parallel noun phrases are found in the two variants, resulting in a pair of sentences that are near paraphrases.

- (20) a. ‘They baked unleavened bread from the dough.’  
b. They baked the dough into unleavened bread.

In one variant the object of the verb denotes the result of baking and the raw material is expressed in a prepositional phrase headed by *from*. In the other variant the object of the verb denotes the raw material and the result is expressed in a prepositional phrase headed by *into*.

## 1.5 The Instrumental Subject Alternation

The verb *bake* may take as an optional adjunct a prepositional phrase, such as *in the oven* in (21a), indicating the device used for baking.<sup>9</sup> This device may also be expressed as the subject of the verb *bake*, giving rise to the INSTRUMENTAL SUBJECT ALTERNATION in (21).

- (21) a. ‘. . . would be to bake it whole in the oven . . .’  
b. ‘This oven bakes magnificent bread.’

---

<sup>9</sup>For reasons that are not well understood, some instruments are marked with a locative preposition (e.g., *in*, *on*) rather than the instrumental preposition *with* when they occur with certain verbs. We consider (21) an instance of the instrumental subject alternation (*The farmer loaded the truck with a crane./The crane loaded the truck.*), since locations are found in subject position much more rarely than instruments.

Example (21b), the only occurrence of an instrumental subject use of *bake* in the BCET corpus, involves the creation sense of *bake*, although we expect that instrumental subjects would also be found with change of state *bake* in a larger corpus (e.g., *the oven baked the potatoes too hard*).

## 1.6 The Instructional Imperative Alteration

*Bake* is also found in a transitivity alternation where the intransitive variant always requires the imperative form of the verb in English. These intransitive constructions are frequently found in cookbooks, do-it-yourself manuals, product labels, and any other texts where directions are being given.

- (22) a. ‘Bake the pastry for ten minutes.’ (Transitive)  
b. ‘Pipe or pile the potato on top of the fish and bake in a pre-set oven for 45-50 minutes . . .’  
(Intransitive)

The INSTRUCTIONAL IMPERATIVE ALTERNATION resembles the indefinite object deletion alternation (Section 1.2 above), since it involves an unexpressed but understood object in the intransitive variant. However, the understood object here is interpreted as a specific object whose identity can be determined from the context (Fillmore 1986). A second difference is that this understood object can be modified by various adjunct phrases, as in the example just given. For more extensive discussion of the properties of this construction see Massam and Roberge (1989).

A cursory examination of the five entries for *bake* surveyed so far shows that a treatment of the verb’s behavior with respect to each of these alternations is lacking. With an overview of a wider range of entries, we could show in equal detail that each of these alternations, like the others considered in previous sections, is either not signalled at all (the instructional imperative alternation) or not signalled consistently or comprehensively (the instrumental subject and material/product alternations). We consider it more profitable to use the remainder of this article to make a constructive proposal about the handling of semantic-syntactic interdependencies.

## 1.7 Summary

We have seen that dictionaries use various cues to alert the user to the presence of semantic-syntactic interactions. We briefly summarize the types of cues:

- Cues in the organization of the senses. (The splitting or lumping of related uses under one or more senses as in the association of the definitions “cook, be cooked” within a single sense category of *bake* in OALD.)
- Cues in the syntactic codes and patterns. (“[I;T]” in LDOCE; “vi, vt” in OALD)
- Cues in the form of the definition. (“to (cause to) cook” in LDOCE)

- Cues in example sentences:

- Use of noun phrases chosen to emphasize the semantic relation of an argument to the verb, as well as to represent the prototypical semantic restrictions on this argument. (“*bake bread/cakes*” in OALD)
- Use of same noun phrases across example sentences to bring out different expressions of arguments. (“*to bake bread/The bread is baking*” in LDOCE)
- Outright translations or annotations intended to convey the intended interpretation. (“**She bakes every Tuesday (makes bread)** elle fait du pain le mardi” in CREFD)
- Use of parentheses around a noun phrase in an example to indicate not only that it is optionally expressed but also that the interpretation of the example does not change when it is absent.<sup>10</sup>
- Use of more complex sentences to provide extra context that biases the dictionary user towards a particular interpretation of a verb form that might not otherwise be available.

We recognize the value of these cues when well-used, but the contribution of these devices to the overall entry is limited by the lexicographer’s understanding of the word under consideration, and by the ability of the learner-user to interpret them correctly.

Each of the points of criticism we have made in this section highlights an instance where *bake* occurs in a pair of syntactic frames that are related by a systematic semantic relationship. The particular relationship may or may not be a paraphrase relation (compare the benefactive alternation with the causative/inchoative alternation). Furthermore, the relationship is never idiosyncratic to *bake* but is to be found with a wide class of verbs of a particular semantic type. For example, the causative/inchoative alternation is characteristic of verbs of change of state, which include not only the cooking verbs noted above but also such verbs as *open*, *break*, and *melt* (see (3)). Similarly, each of the other alternations is associated with specific semantically coherent classes of verbs (see Levin 1985, in press; Rappaport and Levin 1988; Rappaport, Levin, and Laughren 1988).

Lexicographers try to make explicit and consistent every facet of language that they know enough about. The semantic-syntactic interdependencies under review are not adequately treated just because the lexicographer does not have enough systematized knowledge of what is going on to allow this to be done. A theory of lexical organization is needed in order to provide the context for building the entry. And it is the lack of theory in the semantic-syntactic interdependency area that makes these dictionary entries less than adequate.

Let us see how lexicographers can draw on linguistic theory to give them the needed foundation for building more systematic lexical entries.

---

<sup>10</sup>These last two devices are not exemplified here, but are discussed in Atkins, Kegl, and Levin (1986).

## 2 Towards a Better Entry for *Bake*

In order to improve the entry for *bake*, we need to focus on the root of the problem: the existence of generalizations which could have been recognized but were not. Even when provision is made in the design of the dictionary for the handling of alternations (as in the case of the benefactive alternation in SPEJD in Section 1.3 above), these design decisions are not implemented consistently, presumably because the lexicographers are not able to identify all the verbs to which an alternation applies. It appears at the moment that the inclusion of a property in some entries and its omission in others is not the result of a principled decision on the part of the dictionary editor. Such decisions must be made according to rational criteria. We realize that space constraints and the demands of a specific dictionary design operate in this decision-making process; the inclusion of such structures as *Poach me an egg* may not be of enough importance to warrant inclusion in general learners' dictionaries. However, this is not true of all the omissions we have identified: in many cases more minor constructions have been included in the same entry. The lexicographer's response to such constraints should not result in basic inconsistencies of treatment. It is not enough to recognize the existence of the alternations. It is necessary to understand their properties; for instance, which verbs are found in a given alternation and what conditions govern a verb's ability to participate in it. Only with this knowledge can a lexicographer make informed and principled decisions about entry design based on valid criteria.

The reason we have been able to pinpoint all the inadequacies discussed in Section 1 and to diagnose their causes is twofold:

- A range of semantic-syntactic interdependencies in English have been established and related to semantically coherent classes of verbs.
- Studies of lexical organization have identified essential classes of verbs, as well as the central properties characterizing verbs of each type.

In the case of the entries examined in this paper, we were already alerted by the meaning of the verb *bake* to its various syntactic properties, as well as to certain other verbs which, sharing elements of meaning, should pattern syntactically in the same way. Thus we came to our study with an understanding of the properties that the verb *bake* shows and the reason that it shows them. We suggest that theoretically guided investigations into lexical organization can provide the lexicographer with a starting point for a systematic treatment of a lexical item in any type of dictionary.

Although our understanding of this facet of language is still far from complete and the chicken-and-egg relationship between semantics and syntax is hotly debated, the generalizations made about semantic-syntactic interdependencies may already be of value to lexicography. Of course in dictionary entries a word is never approached in isolation; however, areas which have not previously been systematized have been treated as though the properties involved were idiosyncratic to that word. The knowledge that *bake* in one of its uses is a change of state verb allowed us to predict that it would participate in the causative/inchoative alternation; the knowledge that *bake* in one of its uses is a verb of creation allowed us to predict that it would participate in the benefactive alternation. A systematic, comprehensive and comprehensible list of the properties of

a particular verb allows more principled decisions on the part of the lexicographer. A dictionary constructed on the basis of principled decisions is a better dictionary than one which is not: every user appreciates consistency of approach.

On the basis of the increased ability to generalize accorded by a deeper knowledge of lexical organization, one might envisage a process of dictionary-building which would consist of two steps:

- Setting out a summary of the options in what we term a MACRO-ENTRY; this is a lexical entry that makes explicit what a native speaker knows about a word. It represents as faithfully as possible the native speaker's lexical competence with respect to that word. The macro-entry we envisage is dictionary-neutral. It is simply a systematic and exhaustive presentation of facts, informed by a theoretical understanding of lexical organization which ensures that relevant facts are represented.
- Tailoring the information in the macro-entry to the needs of a specific application, i.e., selecting from the options laid out in the macro-entry those that are appropriate for inclusion in a particular dictionary and presenting them in a manner suitable to the users of the intended dictionary.

The existence of such a macro-entry would allow lexicographers writing dictionary entries to devote their lexicographical skills to the optimal presentation of the facts. However, lexicographic skills are also required for the first operation, where the theory needs to be set out in the way which is most helpful to someone writing a dictionary entry. This may not always be the way in which the theorist would choose to expound the theory and we recognize that it may not always seem to do justice to the exceedingly complex theory which informs the description. However, the first priority is that the lexicographer should be offered comprehensive information, clearly and systematically presented.

We have tried to implement these two steps with regard to one of the entries for the verb *bake* which we have studied above. First, with the help of the THEORY, we draw up a macro-entry recording the semantic classes the verb belongs to and the various alternations in which it participates. Then in an attempt to improve the PRACTICE, we use this macro-entry to amend systematically the entry in the CREFD.

## 2.1 Applying Linguistic Theory to Lexicography

When we first approached *bake*, with the help of the BCET Corpus and supplementing our observations by the study of various dictionary entries, we noted that *bake* participated in the causative/inchoative and the benefactive alternations. This observation allowed us to trace back through the theory and recognize that we are dealing with two semantic types: (1) change of state (from the causative/inchoative alternation) and (2) creation (from the benefactive alternation). This fact in turn allowed us to predict other alternations in which *bake* would participate, eg. the indefinite object alternation, which is found with creation verbs.

However, when we looked at the structure of the various dictionary entries, we saw little evidence that the creation sense of *bake* had been identified. We noted that in

most entries the verb had been treated on the assumption that it was simply a change of state verb with various idiosyncratic features, different features being cited in different dictionaries. We realized that a clearer description of *bake* would emerge if the fact that the verb belonged to two semantic classes was used to establish two distinct senses for this word, i.e. the sense of ‘to change the state of something by cooking it in dry heat’ and that of ‘to create something by cooking in dry heat’.

We were pleased to note that in fact the target language in the CREFD entry reflects to a large extent this distinction. The ‘change of state’ sense elicits (*faire*) *cuire* and the ‘creation’ sense produces *faire* in French. We believe that French is a language where the cooking verbs are simply change of state verbs and do not allow a creation sense. This proposal is consistent with other properties that set French apart from English discussed in Talmy (1985) and Green (1973), properties whose implications for lexicography we hope to examine at some future date. This observation about the French entry both supports our claim that the two types of *bake* must be distinguished and shows that the distinctions are useful in the preparation of entries.

A macro-entry should of course show the complete range of alternations associated with the relevant senses of the verb. However, in order to simplify the exposition, we have restricted ourselves to listing only those facts which would be of use to lexicographers writing entries for one-volume general dictionaries of the type we have consulted. We are not trying to cover even this subset of relevant facts in the following description. We restrict ourselves only to those properties discussed in Section 1. A macro-entry for a verb would of course include many other structured descriptions of such dimensions as extended uses (eg. verb + preposition, verb + adverb, verb + adjective forms, etc.), Aktionsart, figurative senses, collocates, idioms, morphologically related words, phonetic and morphological information and so on.<sup>11</sup> We are simply trying to give the part of such a macro-entry which holds the information that could have helped the lexicographers in the specific case of *bake*.

## 2.2 Extract from a Macro-Entry for *bake*

In formulating part of the macro-entry for *bake* we have tried to keep in mind the question of what is the most useful way to present information to the lexicographer, while trying to anticipate the questions a lexicographer would want to have answered before tackling the entry for this verb. We believe these questions would be:

- (syntactic) What are the various word classes the verb belongs to, and what syntactic structures is it found in?
- (semantic) What are its various definable senses (and subsenses)?
- (semantic-syntactic) Which alternations does this verb participate in?

---

<sup>11</sup>We want to comment briefly on the relation between our macro-entries and the entries of Mel'čuk and his colleagues in the Explanatory-Combinatory Dictionary (Apresjan et al. 1970, 1973). This dictionary presents the behavior of each word with respect to a large number of lexical functions that provide an extremely detailed specification of its collocational properties. These functions are very different from the semantic-syntactic interdependencies that we have focused on. Both types of information have an important place in lexical entries.

The table below provides a synthesis of the information about the different uses of *bake* that the lexicographer needs in order to construct an entry in the areas we are considering. The table covers the two senses of *bake* we have identified.

1. (change of state) to cook something in dry heat in an oven
2. (creation) to create something by cooking in dry heat in an oven

Table 1: Summary of Properties of *bake*

Sense	Example	Tr	Link
a 1	I baked the apples.	vt	ii/ci
b 1	The apples were baking.	vi	ci
c 1	Bake for an hour in a hot oven.	vi	ii
d 2	I baked the pastry into a pie.	vt	mp
e 2	The oven baked the apples too hard.	vt	is
f 2	I've baked a pie.	vt	io/is
g 2	I've baked a pie for you.	vt	bf
h 2	I've baked you a pie.	vt	bf
i 2	I always bake on Tuesdays.	vi	io
j 2	I've baked a pie from the pastry.	vt	mp
k 2	This oven bakes good pies.	vt	is

---

Abbreviations: bf—benefactive alternation, ci—causative/inchoative alternation, ii—instructional imperative alternation, io—indefinite object alternation, is—instrumental subject alternation, mp—material/product alternation.

---

In this table we have adopted the same device as the dictionaries we have explored, the use of parallel noun phrases in the examples, to bring out the regularities. However, although we are offering stylized examples to highlight the interrelations between the nsentences, each type shown above, with the exception of (e), is attested in the BCET Corpus:

- (23) a. ‘The women showed me how they baked eggs in clay.’
- b. ‘While the soufflé was baking and she was searching around for something to do . . .’
- c. ‘Bake at gas mark 6 (400 F) for forty-five minutes.’
- d. ‘the new corn, whether raw, roasted, or baked into bread, . . .’
- e. [NO ATTESTED EXAMPLE]
- f. ‘. . . brightly saying she would bake a cake’
- g. ‘Jennifer had baked a special cake for Alexander.’
- h. ‘Bake us plenty of bread, therefore, tonight.’
- i. ‘Thank God I baked this morning. I brought over a nut cake.’

- j. ‘They baked unleavened bread from the dough that they had . . .’
- k. ‘This oven bakes magnificent bread.’

Thus the uses in (a)-(e) represent the ‘change of state’ use of *bake*, while those in (f)-(k) represent the ‘creation’ use. Each use has also been annotated with transitivity (we have subsumed the ditransitive use, (h), under the label *vt*). Finally, the rightmost column labelled ‘Link’ identifies the alternations each use is associated with. For example, use (a) is one of the variants of the causative/inchoative (ci) and the instructional imperative (ii). Looking down this column we see that uses (b) and (c) respectively are the other variants in these alternations. We chose the label ‘Link’ for this column because the alternations link the different uses together, providing a way to organize the various uses of *bake*.<sup>12</sup>

### 3 An Amended CREFD Entry for *bake*

Having set out the various uses of *bake* and mapped out the relations between them, we must now examine the original entry in the CREFD in light of this new knowledge and use the facts which are now at the lexicographer’s disposal to formulate a revised entry for this verb in a collegiate-size bilingual dictionary. The existing entry is:

<b>bake</b> . . . 1 <i>vt</i> (a) ( <i>Culin</i> ) faire cuire au four. <b>she ~s her own bread</b> elle fait son pain elle-même; <b>to ~ a cake</b> faire (cuire) un gâteau; <b>~d apples/potatoes pommes/fpl/pommes de terre au four;</b> <b>~d Alaska omelette norvégienne;</b> <b>~d beans haricots blancs à la sauce tomate;</b> <b>V half.</b>
(b) <i>pottery, bricks . . .</i>
vi (a) [ <i>bread, cakes</i> ] cuire (au four).
(b) <b>she ~s every Tuesday</b> ( <i>makes bread</i> ) elle fait du pain le mardi; ( <i>bakes cakes</i> ) elle fait de la pâtisserie tous les mardis.

Any revision of this entry must take into account what is pragmatically possible in revising a book of this type, that is to say, it must conform to the constraints that the dictionary itself imposes: it must not unduly affect the overall size of the book and it must not lead to a total redesign of the lexicography throughout the dictionary. It must result in a dictionary entry which is not only more comprehensive, but also at least as comprehensible for both types of user, the English and the French native-speaker, as these entries are designed for both.

<sup>12</sup>Note that we have not indicated the links between every use. For instance (g) could be linked to (f) because creation verbs allow a benefactive *for* adjunct. (Note that with change-of-state verbs, a *for* phrase, while possible, does not receive exactly a benefactive interpretation.) Similarly, the material/product variant (j) could be linked to the more basic transitive use in (f). For change-of-state *bake*, we have included both variants of the instrumental subject alternation ((a) and (e)), but for creation *bake* only the actual instrumental subject variant ((k)). Furthermore, we have provisionally included one variant of the material/product alternation under the change of state sense and the other under the creation sense of *bake*, thus linking the two senses together. We suspect that further research into lexical organization will reveal that the actual relation of these two uses to the change of state and creation senses of *bake* is more complex than this table suggests.

The ideal solution might at first sight seem to be to break down the ‘cooking’ sense of the verb into six subcategories, reflecting each of the alternations in which the verb is known to participate. There are many practical objections to this: (i) if every verb were treated in this way it would treble the size of the dictionary; (ii) such a detailed analysis, while reflecting the facts that the linguist has identified about this verb, may be too detailed for the non-specialist user, especially the French native-speaker; (iii) such a total restructuring of the verb entries throughout a dictionary would inevitably entail more than a simple rewrite of each individual entry and would certainly destroy the character of the existing book. It is not our purpose here to attempt to design a new dictionary, simply to see in what way our deeper knowledge of the semantic-syntactic properties shared by this verb allows us to improve its treatment in the context of a general bilingual dictionary. This restraint on our aims also prevents us from addressing the problem of compounds (*baked beans*, *baked Alaska*, etc.) at this point; for the moment, we have left such items within the verb entry.

Another possible solution might involve the use of cross-references to link the two variants of each alternation, but space constraints make it impossible to interlace a 2000-page dictionary text with a complex network of cross-references, which in any case most users find distracting rather than helpful.

It is clear from the contents of the existing entry that the lexicographers were aware of the causative/inchoative alternation, since subcategory 2(a) shows the intransitive of this pair, and of the indefinite object alternation, as the relevant intransitive use is given in subcategory 2(b). In the present entry, however, there is no indication of the existence of the other four alternations which we have established, namely the benefactive, material/product, instrumental subject, and instructional imperative alternations.

We begin the revision with the proviso that the new draft should differentiate at least between the ‘change of state’ and the ‘creation’ senses of the verb, as it is this distinction that lies at the heart of the alternations. In the original entry, the transitive uses of both of these senses are dealt with in subcategory 1(a), though it is clear that the lexicographers were aware of this distinction but unable to clarify it. This is obvious from the selection of an example illustrating the ‘creation’ sense (“*she bakes her own bread*”) following the translation “*faire cuire au four*” which applies only to the ‘change of state’ sense. Clearly, the lexicographers realized that “*faire cuire au four*” was not an adequate catch-all target language equivalent. The second example (“*to bake a cake faire (cuire) un gâteau*”), condensing as it does in the translation both these basic senses (‘*faire cuire un gâteau*’—‘change of state’ and ‘*faire un gâteau*’—‘creation’), indicates the lexicographers’ awareness of this ambiguity but is hardly helpful for the non-specialist user.

A first attempt at a revised entry, encapsulating the known facts about *bake*, leads to the following draft, which is almost twice as long as the original, and therefore eventually unacceptable in the context of the CREFD:

**bake** 1 vt (a) (*cook in oven*) *cakes, meat, vegetables, fruit faire cuire (au four)*. **she —d the potatoes in the microwave oven** elle a fait cuire les pommes de terre au four à micro-ondes; **they —d the flour into bread** ils ont fait du pain avec la farine; **—d potatoes/apples pommes fpl/pommes de terre au four; —d Alaska omelette norvégienne; —d beans haricots blancs à la sauce tomate; V half.**

(b) (*make, create*) *cake, soufflé, pie faire.* *she —s her own bread elle fait son pain elle-même; she —d cakes for the children's party elle a fait des gâteaux pour le goûter de fête des enfants; I've —d you a cake, I've baked a cake for you je t'ai fait un gâteau; they —d bread from the flour ils ont fait du pain avec la farine; microwave ovens don't — good bread les fours à micro-ondes ne font pas du bon pain.*

(c) *pottery, bricks* . . .

2 vi (a) (*be cooked in oven*) *cuire (au four). while the potatoes/apples were baking pendant que les pommes de terre/les pommes cuisaient (au four).*

(b) (*in cookbook*) **bake for 30 minutes in a moderate oven** *faire cuire à four doux pendant 30 minutes.*

(c) *she —s every Tuesday (makes bread)* *elle fait du pain le mardi; (makes cakes) elle fait de la pâtisserie tous les mardis; she —d for the children's party elle a fait des gâteaux pour le goûter de fête des enfants.*

(d) *[pottery, bricks]* . . .

An examination of this revised draft entry shows that, while the basic structuring of the dictionary—the separation of transitive from intransitive uses—has been maintained, the number of subcategories has been increased, to take account of the semantic-syntactic interdependencies set out in the macro-entry table in Section 2. By this means, and by the use of a method already familiar to lexicographers, namely that of inserting carefully worded examples into the appropriate sections of the text, all six alternations are now included:

#### CAUSATIVE/INCHOATIVE ALTERNATION

in 1(a): *she baked the potatoes in the microwave oven*

in 2(a): *while the apples/potatoes were baking*

#### INDEFINITE OBJECT ALTERNATION

in 1(b): *she baked cakes for the children's party*

in 2(c): *she bakes every Tuesday*

#### BENEFACTIVE ALTERNATION

in 1(b): *I've baked you a cake, I've baked a cake for you*

#### MATERIAL/PRODUCT ALTERNATION

in 1(a): *they baked the flour into bread*

in 1(b): *they baked bread from the flour*

#### INSTRUMENTAL SUBJECT ALTERNATION

in 1(a): *she baked the potatoes in the microwave oven*

in 1(b): *microwave ovens don't bake good bread*

#### INSTRUCTIONAL IMPERATIVE ALTERNATION

in 1(a): *she baked the potatoes in the microwave oven*

in 2(b): *bake for 30 minutes in a moderate oven*

This new entry is unfortunately too long for the dictionary in question, and must be abridged despite the inevitable loss of information this entails. There are many tried and tested ways of shortening dictionary entries: one could for example reduce the number of subcategories and pack examples of subtly different uses into the same subcategory. We

reject this stratagem for this entry, as we believe it would simply obscure the distinction between the ‘change of state’ and ‘creation’ senses, which is clearly reflected in the French equivalents—*faire cuire* (*au four*) as opposed to *faire*—and is therefore of prime importance in this bilingual dictionary. We are left with the problem of deciding which information not to include; that is, which alternations to omit in this particular entry. Such a decision is not an easy one to make and used to be left to the lexicographer’s instinct about frequency of occurrence in the language. However, the existence of corpus evidence helps us to select the alternations to include in the shortened form of the entry. A study of the 291 verbal ‘cooking’ sense citations in the BCET Corpus reveals that the straightforward transitive uses predominate, both in the change of state sense (*The women showed me how they baked eggs in clay*) and in the creation sense (*I’m going to bake a cake*). As far as the minor patterning are considered, the following frequencies may be noted:<sup>13</sup>

Table 2: Frequencies of Different Uses of *bake*

INSTRUCTIONAL IMPERATIVE:	intransitive	—	62
INDEFINITE OBJECT:	intransitive	—	6
BENEFACTIVE:	ditransitive	—	1
	‘ <i>bake A for B</i> ’	—	4
CAUSATIVE/INCHOATIVE:	intransitive	—	3
MATERIAL/PRODUCT:	‘ <i>bake A into B</i> ’	—	1
	‘ <i>bake B from A</i> ’	—	2
INSTRUMENTAL SUBJECT:	instr as subject	—	1

On the basis of these figures, we decided to omit from the revised entry the treatment of the instrumental subject and the material/product alternations. This removed three examples: “*microwave ovens don’t bake good bread*,” “*they baked the flour into bread*,” and “*they baked bread from the flour*.” It also rendered unnecessary any reference to microwave ovens (inserted to avoid the unnatural-sounding *ovens (don’t) bake good bread*). However, we decided to keep the first example as “*she baked the potatoes in the oven*,” in order to allow the English native-speaker to encode *in the oven* as *au four* (and not *dans le four*). We also kept the instructional imperative (“*bake for 30 minutes in a moderate oven*”). Although the French also has this type of understood object, the English native-speaker user might be ignorant of the need for the infinitive form of the verb in the French counterpart of this construction, rather than the imperative form found in English.

As a further abridgement, we deleted the two ‘*for + occasion*’ examples (“*she baked cakes for the children’s party*” in 1(b), and “*she baked for the children’s party*” in 2(c)). Our reasoning was that English *for* and French *pour* are as it were default equivalents in the two languages, and as such would not constitute any difficulty in either the encoding

<sup>13</sup>The instructional imperative figure is distorted by the fact that cookbook jargon is unduly prominent in this corpus, containing as it does one straightforward cookery book and a book on diets. These books are responsible for 58 of the 62 citations. Nevertheless, we feel that this intransitive use is important enough to be noted in a general dictionary both for decoding and encoding purposes.

or decoding process. We did however maintain the two versions of the benefactive alternation ("I've baked you a cake" and "I've baked a cake for you") because in such pronominal uses the preposition *for* is NOT *pour* in French, but must be translated by a dative pronoun.

We decided to leave the apparently redundant "*she bakes her own bread*" in 1(b), because this use with *own* presents peculiar problems of translation in French (not \**son propre pain*), and as it occurs three times in the corpus it could be needed by the dictionary user looking up *bake*. Another apparently redundant example, "*while the apples/potatoes were baking*" in 2(a), is retained in order to indicate to the English speaker the exact usage in English that this subcategory is treating. This function was fulfilled in the original dictionary entry by the inclusion in 2(a) of typical noun complements ('[bread, cakes]'), but the metalanguage paraphrase which we have substituted for this, together with a simple example, would we believe be of more immediate help to the user. To add square-bracketed noun complements to a round-bracketed paraphrase would lose in elegance what might be gained in economy.

Finally, we simplified some of the original examples, removing the alternatives of the type "*apples/potatoes*" or "*pommes/pommes de terre*", as this information was already being carried by the enriched list of noun complements of the verb headword ("cakes, meat, vegetables, fruit" in 1(a) and "cake, soufflé, pie" in 1(b)).

The following version of the 'cooking' sense portion of the revised entry for *bake* results from the process of abridgement:

**bake 1 vt (a) (cook in oven) cakes, meat, vegetables, fruit faire cuire (au four). she —d the potatoes in the oven elle a fait cuire les pommes (fpl) de terre au four; —d potatoes pommes de terre au four; —d Alaska omelette norvégienne; —d beans haricots blancs à la sauce tomate; V half.**

**(b) (make, create) cake, soufflé, pie faire. she —s her own bread elle fait son pain elle-même; I've —d you a cake, I've —d a cake for you je t'ai fait un gâteau.**

**2 vi (a) (be cooked in oven) cuire (au four). while the apples were baking pendant que les pommes cuisaien(t) (au four).**

**(b) (in cookbook) — for 30 minutes in a moderate oven faire cuire à four doux pendant 30 minutes.**

**(c) she —s every Tuesday (makes bread) elle fait du pain le mardi; (makes cakes) elle fait de la pâtisserie tous les mardis.**

Even in this abridged form, it may be seen how much the original entry has benefited from a systematic check against the full range of the verb *bake*'s semantic-syntactic relationships. The revised entry is undoubtedly longer than the original; within the 'cooking' sense there are five subcategories including nine examples in the revised entry compared with the original's three subcategories and six examples. On the other hand, the revision certainly offers a clearer exposition of the semantic and syntactic properties of the verb and considerably more guidance to the dictionary user of either native language on how the verb functions and on what the French equivalents are.

## 4 Conclusion

By studying existing dictionary entries, we have shown that facts which pertain to a system within the language are not being dealt with in a systematic way, and we have suggested that this lack of system where system exists leads to an impoverished lexicography. In revising a dictionary entry which was compiled by one of us before she became aware of the systematic semantic-syntactic relationships involved, we have tried to show that when the potential of a verb is fully known it is possible even within the constraints of a one-volume general dictionary to give this verb more comprehensive coverage, without making the entry so complicated as to be unintelligible to all but the hardened semanticist and syntactician. In our view this revised entry does more than that: it reflects generalizations in English, and as such its structure may be repeated at other verb entries where the same generalizations obtain.

We would argue that if dictionaries are to take a quantum leap rather than simply a series of tottering steps into the future—particularly into a future which holds out the enticing prospect of electronic reference works—then what is necessary is not merely a matter of elaborating or modifying existing entries, but rather an almost total restructuring of the way in which verbs are treated in lexicography. Any given verb participates in only a subset of the possible alternation patterns. However, if a dictionary is to provide comprehensive and consistent information about alternations, these must at least be available to the dictionary designer and to the lexicographer during the process of compilation. The entries of verbs that do not take up certain options must parallel the entries of those that do.

As a prerequisite to such a restructuring, the range of complex relations and interdependencies between word senses, definitions, syntactic coding, and examples must first be spelled out. Making the implicit knowledge encoded in a dictionary explicit is possible only in the context of a theory of lexical organization. Linguists can contribute to work in lexicography by providing such a theory. Theoretically guided linguistic investigations into lexical organization can provide the lexicographer with a starting point. We were quickly able to identify the inadequacies and omissions of the type described in this paper by drawing on our studies of lexical organization. Through these studies (Levin 1993), we have identified essential classes of verbs as well as the central properties characterizing verbs of each type. Once the classes manifesting each relation is identified, for example in this instance change of state verbs or creation verbs, it becomes possible to check that the entries of its members are treated uniformly with respect to the encoding of various alternations, by ensuring that they share the same syntactic coding, type of definition wording, and patterning of examples. The construction of dictionaries should take advantage of, and could be eased by, theoretical linguistic work on the organization of the lexicon.

This Journal, by bringing together as it is designed to do, the work of the theoretical linguist and that of the practicing lexicographer, is, we believe, a good place to repeat our plea for more such collaboration (Atkins, Kegl, Levin 1986). Lexicography is a craft—a good lexicographer should be able to write a dictionary to almost any specification, and as long as it is carefully and consistently written and based on a good theoretical description, it will be a good dictionary of its kind. The peculiar skills of a lexicographer are reductive skills—to take the facts and reduce them to something useful to a specific

group of dictionary users. The theoretical linguist, on the other hand, who attempts to identify the innate order of the entire language, brings quite different skills to bear on the problems. What the lexicographer has to offer the linguist is the ability to provide any number of examples, counter-examples and related phenomena, and the discipline inherent in a task which starts at *A* and arrives at *Z* within a fixed period of time, without missing out any relevant linguistic fact, however recalcitrant. What the linguist has to offer the lexicographer is a clearer route map. In our experience, collaboration between linguist and lexicographer is enriching for both and is indeed essential if the language is to be clearly and comprehensively analyzed to meet the more sophisticated needs of the years to come.

### Acknowledgments

We are happy to dedicate this joint paper to the memory of Don Walker, who first encouraged us to work together. We have enjoyed and benefited from this collaboration, thus confirming Don's belief, and now ours, that linguistics and lexicography have much to offer each other. This paper develops a theme first presented in an earlier joint paper, Atkins, Kegl, and Levin (1986). This earlier paper had its roots in discussions the three of us began at the Automating the Lexicon Workshop held in Grosseto, Italy in May 1986, a workshop that Don was instrumental in organizing.

Kegl's work was supported in part by a grant to Princeton University from the James S. McDonnell Foundation. Levin's work was supported by a grant to the Lexicon Project of the MIT Center for Cognitive Science from the System Development Foundation. Atkins, in her then capacity as General Editor of the *Collins-Robert English-French Dictionary*, was responsible for the English source language text in that dictionary.

We would like to thank Hélène Lewis, editor of the *Collins-Robert English-French Dictionary*, for her help in constructing the amended *bake* entries in Section 3.

The corpus citations in this paper are taken from the Birmingham Collection of English Text, including the COBUILD Corpus, which is supported by funds from Collins Publishers and held at the University of Birmingham, England.

## Appendix: Corpus Frequencies for *bake*

The tables below summarize the frequency of occurrence of the verb *bake* in the BCET Corpus (approximately 17.8 million words). First consider the statistics for the number of citations for each morphologically distinct forms of *bake*:

Citations for:	<i>bake</i>	100
	<i>baked</i>	208
	<i>bakes</i>	4
	<i>baking</i>	121
Total citations:		433

These 433 citations can be subdivided according to the type of use of *bake* involved:

Non-'cooking' senses:	44
Compound uses:	
<i>baked beans</i>	33
<i>baked potatoes</i>	23
<i>baking dish</i>	8
<i>baking pan</i>	2
<i>baking paper</i>	1
<i>baking powder</i>	5
<i>baking sheet</i>	9
<i>baking soda</i>	8
<i>baking tin</i>	7
<i>baking tray</i>	2
Total compound uses:	98
'Cooking' sense:	291

We see that out of a total of 433 citations for *bake*, 142 citations are for either the non-'cooking' sense or the compound uses of this verb, leaving a total of 291 citations for the 'cooking' sense of *bake*. See Section 3 for a breakdown of the citations for the 'cooking' sense of *bake* in terms of the various alternations described in this paper.

## References

- [1] Apresjan, Ju.D., I.A. Mel'čuk and A.K. Žolkovskij. 1970. 'Semantics and Lexicography: Towards a New Type of Unilingual Dictionary' in F. Kiefer (ed.). *Studies in Syntax and Semantics*. Dordrecht: Reidel.
- [2] Apresjan, Ju.D., I.A. Mel'čuk and A.K. Žolkovskij. 1973. 'Materials for an Explanatory Combinatory Dictionary of Modern Russian' in F. Kiefer (ed.). *Trends in Soviet Theoretical Linguistics*. Dordrecht: Reidel.
- [3] Atkins, B.T., A. Duval, R.M. Milne, et al. (eds.). 1987. *Collins-Robert English-French Dictionary*. London and Glasgow: Collins Publishers. Second Edition. (CREFD)
- [4] Atkins, B.T., J. Kegl, and B. Levin. 1986. 'Explicit and Implicit Information in Dictionaries.' Lexicon Project Working Papers 12. Center for Cognitive Science, MIT, Cambridge, MA. Also Cognitive Science Laboratory Report 5. Cognitive Science Laboratory, Princeton University, Princeton, NJ.
- [5] Fillmore, C.J. 1986. 'Pragmatically Controlled Zero Anaphora.' *Proceedings of the Berkeley Linguistics Society* 12. Berkeley, CA, 95-107.
- [6] Green, G. 1973. 'A Syntactic Syncretism in English and French' in B. Kachru et al. (eds.). *Issues in Linguistics*. Urbana, IL: University of Illinois Press.
- [7] Hale, K.L. and S.J. Keyser. 1986. 'Some Transitivity Alternations in English.' Lexicon Project Working Papers 7. Center for Cognitive Science, MIT, Cambridge, MA.
- [8] Hale, K.L. and S.J. Keyser. 1987. 'A View from the Middle.' Lexicon Project Working Papers 10. Center for Cognitive Science, MIT, Cambridge, MA.
- [9] Hanks, P., et al. (eds.). 1986. *Collins English Dictionary*. London and Glasgow: Collins Publishers. Second Edition.
- [10] Hornby, A.S. (ed.). 1974. *Oxford Advanced Learner's Dictionary of Current English*. Oxford: Oxford University Press. (OALD)
- [11] Konishi, T., et al. (eds.). 1980. *Shogakukan Progressive English-Japanese Dictionary*. Tokyo: Shogakukan. (SPEJD)
- [12] Levin, B. 1985., 'Lexical Semantics in Review: an Introduction' in B. Levin (ed.). *Lexical Semantics in Review*. Lexicon Project Working Papers 1, Center for Cognitive Science, MIT, Cambridge, MA.
- [13] Levin, B. In press. 'Approaches to Lexical Semantic Representation' in D. Walker, A. Zampolli, and N. Calzolari (eds.). *Automating the Lexicon*. Oxford: Oxford University Press.
- [14] Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

- [15] Massam, D. and Y. Roberge. 1987. 'Recipe Context Null Objects in English.' *Linguistic Inquiry* 20, 134-139.
- [16] Procter, P., et al. (eds.). 1978, revised 1987. *Longman Dictionary of Contemporary English*. London: Longman Group. (LDOCE)
- [17] Rappaport, M. and B. Levin. 1988. 'What to Do with Theta-Roles,' in W. Wilkins (ed.). *Syntax and Semantics 21: Thematic Relations*. New York: Academic Press, 7-36.
- [18] Rappaport, M., B. Levin, and M. Laughren. 1988. 'Niveaux de Representation Lexicale.' *Lexique* 7, 13-32. Appears in English as 'Levels of Lexical Representation,' in J. Pustejovsky (ed.). 1993. *Semantics and the Lexicon*. Dordrecht: Kluwer, 37-54.
- [19] Sinclair, J., P. Hanks, et al. (eds.). 1987. *Collins COBUILD English Language Dictionary*. London and Glasgow: Collins. (COBUILD)
- [20] Talmy, L. 1985. 'Lexicalization Patterns: Semantic Structure in Lexical Forms' in T. Shopen (ed.). *Language Typology and Syntactic Description 3. Grammatical Categories and the Lexicon*. Cambridge: Cambridge University Press.

# Issues for Lexicon Building\*

Nicoletta Calzolari  
Istituto di Linguistica Computazionale del CNR, Pisa  
*e-mail:* [glottolo@vm.cnuce.cnr.it](mailto:glottolo@vm.cnuce.cnr.it)

## Abstract

Taking for granted the importance of a lexical component in any linguistic application within the so-called ‘language industry’, this paper aims at highlighting some of the issues which have to be carefully evaluated with regard to any future actions for building large-size computational lexicons. The following aspects, considered of crucial relevance, particularly if one aims at a set of interconnected lexicons for different languages, are presented: design of lexicon architecture, role of user needs, need for standardization, multilinguality, role of textual corpora, organizational aspects, intellectual property rights, and the role of national and international bodies.

## 1 Role of the Lexicon in Language Engineering

It is almost a tautology to affirm that a good computational lexicon is an essential component of any linguistic application within the so-called ‘language industry’, ranging from NLP systems to lexicographic enterprises. Wherever words are used, a computational lexicon is required, obviously with very different requirements regarding size, depth of information, coding systems, accuracy, type of formalism, etc., in accordance with different applications.

Given the increasing importance and economic impact of the language industry worldwide, and, on the other hand, the great variety of different systems and applications which are subsumed under the term ‘language engineering’, the pressing question which faces us is the following: do we have to continue in the traditional way of developing new—and different—lexicons for any new application/system, starting from scratch every time and therefore consuming time, money and manpower, or is it timely to think of the possibility of making the effort to converge, trying to avoid unnecessary duplications and—where possible—building on what already exists? We claim, and will argue, that the second direction is the best alternative. Some possible strategies leading to this convergence of efforts will be outlined.

---

\*This is a revised version of a “strategic paper” produced within the MLAP (Multilingual Action Plan) project NERC (Network of European Reference Corpora) funded by the EC, DG XIII and coordinated by the Consorzio Pisa Ricerche (A. Zampolli). The partners of the NERC project were INALF-Paris (France), Universidad de Malaga (Spain), INL-Lieden (The Netherlands), IDS-Mannheim (Germany), University of Birmingham (U.K.), Istituto di Linguistica del CNR and Università di Pisa (Italy).

This must be seen as a rather generic discussion paper, almost in the form of a list of points, aiming simply at highlighting some of the issues to be carefully evaluated, without going into any details or depth of argumentation.

## **1.1 Machine Readable Dictionaries vs. Computational Lexicons**

At the outset, it is useful to make a terminological distinction: by ‘machine-readable dictionary’ (MRD) we mean the electronic version of traditional printed dictionaries, usually prepared for typesetting purposes, while by ‘computational lexicon’ (CL) we mean any electronic dictionary not explicitly prepared only for the purpose of building a paper version of it (for a more detailed distinction see Calzolari 1989). This distinction does not exclude the fact that MRD’s can be a very valuable input in the process of building CL’s.

In this paper we concentrate on the properties of CL’s; MRD’s will be considered only as one possible way of acquiring data to populate the CL we aim for, and hence the issue of the feasibility and cost-effectiveness of processing and reusing MRD’s will be pointed out.

## **1.2 Strategic Relevant Issues**

A number of issues are of crucial relevance in the design of a strategic policy for lexicon building, namely:

- lexicon architecture
- user needs
- standardization
- multilinguality
- role of textual corpora
- organizational aspects
- IPR (intellectual property rights)
- the role of national and international bodies.

All these aspects must be carefully evaluated in any decision for future action in the lexical area.

In this paper, these aspects will be concisely presented, mostly in the form of lists of relevant points to be taken into account within each aspect, sometimes pointing at possible alternatives, sometimes forming suggestions.

## 2 Lexicon Architecture

The issue of the overall architecture of the lexicon involves a number of parameters (presented below) which are strongly correlated with the objective of the lexicon itself, and therefore with the points outlined in section 3 of this paper.

As a simple example we can mention the obviously very different requirements of a spelling checker vs. a specialized domain machine translation system with respect to lexicon architecture. The former requires a very extensive but very flat lexicon, while the latter requires a smaller but very detailed, well articulated and complex dictionary.

If we aim at a reusable multifunctional lexicon, obviously our choices, for each parameter (such as size, depth of information, levels of linguistic description, explicitness, etc.), should be guided by the most demanding, the widest, and the deepest requirements. However, these should be corrected by a careful analysis of a) the possibility of actually achieving these most general requirements, in accordance with the state-of-the-art (in linguistic theory, software technology, etc.), and b) the cost-effectiveness of such ambitious prospects as opposed to more modest but more realistic and economically viable strategies.

A balanced choice could lead to an intercept strategy, where long-term actions are planned to build on the results of short-term actions. The need of achieving "reusable" short and medium term results is therefore essential.

### 2.1 Size

Meeting the need of extensive and almost exhaustive size is the easiest to attain. The input for data are MRD's, existing CL's and, obviously, text corpora. Provision should be made, by means of appropriate coding, for the possibility of extracting different types of sub-lexicons most suited to specific purposes and applications. A study should therefore be conducted for different requirements in terms of lexicon size, and consequently lexicon composition, originating from various applications.

These requirements should be properly coded to allow extraction and definition of specific sub-lexicons, in particular for specific domains. For this purpose it is important to be systematic in taking into account, and properly coding (possibly in a standardized way), a relevant number of sociolinguistic levels and a large range of sublanguages.

Other related issues are maintenance and updating, both with respect to actual words and to their codes (for neologisms, words becoming obsolete, specialized words becoming of general use, etc.).

Another issue of relevance is the definition of 'lexical unit': are multi-words to be considered independently as lexical units? If so, which types? Probably, all types where the meaning of the whole is different from the sum of its parts. The basic Lexical Unit will then be of a variable type and some vagueness may have to be accepted. A study on collocations and their typology is therefore important (see Heylen, 1993).

### 2.2 Depth of Information

Complementary to size, or as a different perspective on the size issue, is the issue of "depth" of information attached to each lexical unit. This is very much dependent

on parameters such as a) type of application, b) level of linguistic description, and c) state-of-the-art in linguistic theory.

- (a) This parameter is rather simple: there are applications requiring almost no information attached to a lexical entry (e.g. simple spelling checkers), while other applications require “everything”, including what is not yet available in current theories, such as extended world-knowledge (e.g. machine translation).<sup>1</sup>
- (b) and (c) These two parameters are strictly interconnected. The situation is very different at the various levels of linguistic description:
  - i. Lower levels of linguistic description (orthography, phonology, morphology, morphosyntax) are more easily and thoroughly defined, a consensus and a standardization can be more quickly achieved, and a very detailed level of description can be obtained without much effort. Quite a number of proposals already exist to be considered as a sound basis for development (e.g., those coming from Multilex, Genelex, NERC, EAGLES), and data exist for many European languages.
  - ii. At the level of syntactic description there is a well-motivated hope (see ET-7 final report, Heid and McNaught 1991) of reaching a consensus on the standardization of at least a core set of syntactic features with a well-defined type of short-term action (e.g. the one being proposed in EAGLES). Also on this level one can build on a number of existing proposals (e.g., Multilex, Genelex, Eurotra), without omitting proper consideration of the more prominent theoretical linguistic approaches, and also on recent corpus-based approaches (not only statistical ones).
  - iii. At the higher levels of “deep syntax”, semantics, pragmatics and world-knowledge, the possibility of obtaining a short-term consensus, standardization, and, therefore, complete reusability is more difficult. Less data from which to start exist, and less stable achievements from theoretical linguistics are available. These areas still deserve research, but a) feasibility studies should be undertaken (as in the programme of EAGLES) to define short- vs. medium- vs. long-term actions, b) proper planning should be made for short-term actions, c) some achievements should be evaluated (e.g., those obtained in Acquilex and DELIS), and d) coordination of research and development towards a stepwise strategy should be promoted (involving the most relevant actors in these areas).

### **2.3 Explicitness and Formalization**

Explicitness is, in principle, such an obvious requirement, at least for an NLP lexicon, that it needs little argument in support of it. Again, explicitness can be obtained on all levels of description, but it is certainly more difficult at the semantic level, where, at an intermediate stage, and for human consumption, natural language definitions can

---

<sup>1</sup>A table summarizing different requirements of a number of applications, taken from the ET-7 report on Lexical Applications (McNaught et al., 1990), is appended as Figure 1.

be foreseen as one type of non-explicit information. The same is true for “examples”, possibly derived from text corpora, to exemplify actual usages in different constructions, meanings, etc.

This issue is connected with the necessity of devising a set of explicit criteria and tests for classification (for each feature and category) which are absolutely necessary if the aim is a reusable lexicon, or even if the aim is a number of reusable, shareable, compatible and integratable lexicons, to be built in an incremental way, by different groups, for different languages.

Two dimensions need to be taken into consideration when thinking of the choice of a formalism:

- i. Implicitness/Inheritance It may not be possible or desirable to spell out in detail every feature of a lexical unit. Much can follow automatically by inheritance.
- ii. Imprecision and vagueness Natural languages are not formal languages, and very often the properties of their words cannot be made fully precise. Probably the optimal representation for word-meanings will be composed of conflicting and interacting ‘criteria’ or ‘associations’.

The net result of this is that we cannot assume we already know an optimal format or mechanism to at least represent lexical semantics. This must be remembered in any proposal for standards (albeit practical ones).

In the foreseeable future, lexicons of any general utility will have to make provisions for imprecise data fields or data of a preferential nature. The distinction between formal and natural language, as descriptive language or metalanguage, and the need for only the first one in a Computational Lexicon, may not be so clear-cut and easy to achieve. A move towards standardization may well proceed in parallel with the move towards precision and towards comprehensiveness, without losing most of the expressive power of the natural language itself.

## 2.4 General vs. Specialized Language

There are alternative views on this issue. Some claim that specialized language lexicons are a priority because the applications in which they are of use, and the lexicons themselves, are somehow simpler and therefore reachable in a shorter time and with less effort. But it is also true that specialized lexicons, which are being continuously modified and which should be enriched every day, need to be continuously collected, analyzed, and updated, and a significant effort is required for their improvement and maintenance. Neither can specialized domain applications, however, do without the core general lexicon, to be of any use. The two aspects, general vs. specialized languages, are therefore intermingled and borderlines are not clear to define.

With respect to this issue, the matter is more one of priority in the choice of actual lexicon development, not of actual dichotomy. A possible strategy could devise the development of a core general lexicon, to be augmented by a constellation of specialized-language lexicons built according to emerging needs or to the possibility of reusing existing data.

Specialized lexicons might be collected from specialized corpora, and one can claim that specialized language corpora are of great help to people working both on general

language and on LSP. In fact, there is a large amount of general language contained in specialized textual data, whereas the reverse is not true.

'Terminologies' might claim to be the extreme case of specialized dictionaries, which might possibly do without the core general lexicon. Arguably, the definitions of terminologies can be adequately supplied by translation equivalents—and this might be a place where standards could get started, without constraining future development harmfully.

### 3 User Needs

In this framework, a distinction should be made between human users and procedural/machine users (i.e., NLP systems). Both types of users should be considered in the overall strategy aiming at building large European reusable computational lexicons.

#### 3.1 Requirements of NLP Systems

Actual requirements of different systems and applications may be very different (as already pointed out in section 2.2), but a nucleus of types of information most commonly and widely needed can be evidenced. (For a synopsis see Figure 1). It is on this core type of information that efforts should be concentrated. In this core almost all of the information which can be foreseen at the various levels of linguistic description is actually present, but this is of little help. Again it is more a matter of prioritizing among different tasks, taking into account not only the parameters mentioned at the beginning of section 2.2, but also the economic relevance of the different applications, the feasibility of reaching good results with relatively little effort, etc.

#### 3.2 Lexicographic Application/Human Consumption

The following types of lexicographic applications are easy to foresee as relevant for the immediate future, being already products available on the market:

- (a) Publishers will find it always more useful and convenient for them to create lexicographic databases (large repositories of lexical data including citation material and prelexicographic data), from which they can produce different types of printed or electronic dictionaries.
- (b) There will certainly be, in the very near future, a market for electronic dictionaries for human consumption, offering non-traditional features. These will require, apart from the information contained in traditional printed dictionaries:
  - i. that the information be very well structured and that, therefore, different types of information be explicitly tagged in some type of coding system (e.g., in SGML and possibly in a TEI conformant way);
  - ii. that a very good access system be available, with the possibility of extracting data in many possible ways and of creating links among data which are not observed on paper;

- iii. that multimedia systems are considered for more innovative lexicographic products.
- (c) Electronic dictionaries, thesauruses, synonym dictionaries, etc., are already, and will increasingly become, an integral part of many word-processors meant for non-automatic, human use.
- These types of dictionaries will not be more structured or richer than normal dictionaries, and perhaps not very different from those above. Printed dictionaries may be a good source of information for these sorts of electronic dictionaries.

## 4 Need for Standardization

Standardization is one of the more critical issues (not only with regard to the lexicon), and one to which much attention is being directed by different communities—linguists, NLP developers, lexicographers, system engineers, etc.—who may have different needs and immediate goals, but not necessarily diverging ones. There exist different facets of the standardization issue with respect to:

- applications
- linguistic theories/schools
- levels of linguistic description
- formalisms of representation
- data exchange
- evaluation with respect to standards.

### 4.1 Multifunctional vs. Application Oriented

When we consider the variety of applications and systems requiring a computational lexicon as a component, the obvious question to ask is:

- i. whether we should design and develop as many lexicons as exist applications (as has been the trend so far), or
- ii. whether we should aim for the design and development of a multifunctional core lexicon (to be augmented by special sublexicons or specific types of information for specific applications) to be translated via appropriate interpreters and compilers into the application specific lexicons.

A possible architecture for such a multifunctional lexicon was proposed within the ET-7 feasibility study. Also, Multilex and Genelex architectures have been recently evaluated within the EAGLES project, and a preliminary common proposal has been put forward (Menon and Modiano, 1993).

Apart from the feasibility of such an approach, we should evaluate the cost (in time, money and manpower), and the implications involved in the eventual choice of

this model with respect to organizational structure and choice of alternative strategies (central vs distributed development and maintenance of the lexical data, priorities, long-term perspectives, etc.).

## 4.2 Standardization of Lexical Data

**Feasibility with respect to levels of linguistic description** The core aim of the ET-7 feasibility study (EC funded), whose final report should be the starting point for any action to be taken in the area, was to determine the feasibility of achieving a standardization of descriptive linguistic specifications.

It became evident, from the results of the study, that a distinction must be made between:

- i. those levels of linguistic description (orthography, phonology, morphology, morphyyntax, and syntax—at least surface syntax) for which the most recent developments (in linguistic theories/schools and in actual practice in NLP systems) show the possibility of a convergence and a consensus around a core set of phenomena and of basic descriptive devices that allow for the realistic achievement of standardization and that require only a short-term effort and that also have a high probability of being accepted by the NLP community, and
- ii. those levels, in particular semantics, which clearly require more basic research before reaching some signs of a consensus.

**Development vs. research** The above differences, inasmuch as we can approach a level where standardization is technically feasible, should distinguish between those linguistic areas for which we can ascertain rapid development vs. those areas still in need of basic research. The first ones can be considered for short-term extensive lexicon building.

**Methodology of work with regard to different linguistic theories/schools** The ET-7 study proposed, following the lines of the precursor “polytheoretical Pisa group”, a methodology of work pointing towards the identification of the minimal observable facts underlying different generalizations or complex devices used in various schools/theories/systems. This methodological approach should be carefully taken into consideration when trying to develop new proposals for standards, and is in fact at the basis of different EAGLES Working Groups (see e.g., the Lexicon Workplan, 1993).

This methodology should allow for the maximum of theory-neutral and school-neutral approach, which, in the case of a huge general Computational Lexicon, is a key issue. On top of this basic lexicon there may have to be more than one style or format of lexicon apparent to different users: an abstract layer of translation between one style/format and another will be required.

**Ongoing projects/initiatives** There have been, and exist, a number of international lexical projects and initiatives which are already involved in the development of a set of proposals for some type of standardization, such as Multilex, Genelex, Acquilex I and II, Eurotra, Nerc, ET-7, ET-10, TEI, LRE, Delis and EAGLES.

These ongoing projects can provide a valuable input for any large, European, lexicon-building initiative, where it is clearly beneficial to integrate the best of the above.

### **4.3 Representation Issue**

Also in the representation area ET-7, together with ET-6, and more recently the EAGLES Formalisms Working Group (see the Formalisms Workplan, 1993), showed signs of providing an answer to the feasibility question. There seems to be a convergence and a growing consensus around unification-based families of representation formalisms. So far, the implementation of an appropriate formalism seems to be considered as a short/medium term goal.

The proposed formal frameworks should be compared and evaluated through their use in actual applications.

If/when a consensus is reached, it may be advisable to keep the technical, formal, representation platforms stable for a while in order to allow for a uniform and consistent development of the “contents” side and a better evaluation of it.

**SGML and the TEI** Two main scenarios can be conceived of when speaking of “reusability” of lexical (and textual) resources: i) the “data pool” model and ii) the “exchange” model. These two models are not contraposed, but rather complement each other. The first model will be examined in section 7.5 on distribution; the second one does not require a large investment for the creation of large repositories of data, but simply attempts to make possible an actual exchange of existing or newly created data. This model presupposes the definition of a common exchange format, with different possible degrees of levels of standardization (e.g., the syntax of the coding system, or the semantics of the coded data).

Within this exchange model, a proposal has been put forward by the Text Encoding Initiative (TEI which is EC supported), and carried out both by the ET-7 study and other projects, of using SGML as the representation language. This proposal must be taken into account for the “exchange” level of data representation.

## **5 Multilinguality**

An obvious requirement for an EC funded/supported lexicon building project is multilinguality. Not much needs to be stated on this topic. Languages of smaller countries should also be given proper consideration. This may, however, result in increased costs, since it is well known that more data from which to begin exist especially for a restricted number of languages (above all English).

## **6 Role of Textual Corpora**

One of the obvious sources of both lexical data and lexical information of different types are textual corpora. Their exploitation is, however, not at all automatic given the state-of-the-art, and hence manpower with solid competence both in computational linguistics and lexicography is needed.

The links between lexical projects and textual projects must also be stressed for other reasons; one area being the necessity of correlation of lexical categories with tags for corpus annotation. The identification of lexical categories will be heavily dependent

on a study of the environment of occurrences in texts. The classification of lemmata (or of particular inflected word-forms) by sense relies heavily on corpus evidence, and corpus analysis, in turn, is essential in the discovery and definition of contextual clues for assignment of tags.

## 7 Organisational Aspects

### 7.1 Availability and Reuse of Existing Data

The following sources of valuable data for populating a set of European Computational Lexicons should be considered as candidates:

- MRD's (with semi-automatic tools)
- text corpora (with semi-automatic tools)
- existing computational lexicons for NLP systems
- encyclopedias.

A certain number of terminologies and of terminological databases already exist, which are quite helpful to specialists, lexicographers, translators, etc. (additional lexicographically oriented bases also exist for LSP studies). It seems natural to reuse and to merge these databases. Is it possible to imagine a kind of huge, super-database taking into account part of the existing terminological and lexicographical material?

Surveys of what exists in these fields have been recently conducted, and partial evaluation of the data collected has been done.

In any case, what can be extracted from the above sources will constitute only a part of what a CL should contain. The choice of an appropriate type of Institution, where all the necessary types of competence are properly represented, is therefore crucial (see next section).

### 7.2 Professional Profile of Lexicon Builders

One of the aspects to be outlined in the design of a computational lexicon-building project is the professional profile of the "lexicon builders".

A preferred profile would be one in which the typical competence and know-how of a computational linguist, a formal linguist, a lexicographer, and a corpus analyst could be merged and integrated, possibly including some acquaintance with information science. Since such a profile is at present rarely found, if it exists at all, it should be possible to create this mixture of competences within specific university courses. In the meantime, the different types of expertise can be brought together either by making use of those institutions where a distribution of these competences can be found, or by constructing teams composed of persons with these varying backgrounds.

### **7.3 Cost in Money and Time**

The cost to a large multilingual lexical enterprise will vary greatly depending upon which of the many parameters are taken into account and which are to be privileged above the others.

A useful criterion for cost evaluation could be the combination of the following (in sequence): a) the definition of a minimal size, minimal level of encoding, minimal type of information per lexical entry, etc., in order for a lexicon to be considered “reusable”, at least for some applications, and b) calculate the cost of such a “minimal level lexicon” for a given language, cost being the starting point for the evaluation of any lexicon project.

### **7.4 Short- vs. Medium- vs. Long-Term Actions**

Having given careful consideration to all the points above, one should then define:

1. the different steps required for a possible lexicon building procedure, each step composed of many tasks;
2. priorities among the various tasks and the interrelationship between them;
3. a clear subdivision among short-term, medium-term, and long-term actions.

Any realistic procedure should begin with short-term goals; however, one should have a fairly distinct idea of further objectives, so that the first phases become the building blocks on which other layers can afterwards be formed.

### **7.5 Means of Distribution and Access**

Different types of organization can be conceived of. We have already mentioned the “exchange” model vs the “data pool” model, and have given some details on the first, most simple one.

The “data pool” model can, in its turn, be implemented on the basis of different organizational models, among which we can mention at least: i) a centralized model, envisioning a huge multilingual lexicon to be built in a single center by a large team, and ii) a distributed model with a network of lexicon centers, possibly subdivided by nationality. The second model seems the most viable.

For each mode, different types of access and distribution of the data can be envisioned.

## **8 Copyright Problems**

It is crucial that the copyright issue be solved for a “reusable” lexicon. This point must be dealt with in a technical manner, and hence a study could be delegated to legal experts (the same is being done for text corpora within the NERC project).

This point cannot, nevertheless, be solved if the whole lexical area, at least in its essential core, is not conceived at the level of infrastructure, as being of absolute and

primary importance for the development of the language industry. Because the lexicon is considered such an expensive infrastructure, it must be built by means of a large cooperative effort aimed at public utility and consumption.

## **9 Role of National and International Bodies**

International bodies may help to propagate to national bodies the best practice of other national bodies.

Representation in extenso of national languages (and provision of resources) is a job for relevant nations (although in some cases a number of nations could cooperate on one language—e.g. French); but not all nations take equally good care of their languages.

There probably need to be data-repositories/exchange-schemes on the national and the international levels.

**Figure 1: Lexical information required by applications**

	ph	sy	st	du	av	or	mo	ms	ps	sc	ar	th	sr	sf	lr	co	id	tr	su	rh	us	pr	fr	di	dt
sr	H	H	H	H	H		H	M	h	h	h	h	h	h	h	h	h	h	h	h	h	H	h		
ss	H	H	H	H	H		H	M	h	h	h	h	h	h	h	l		h	h	h	h	H	h		
in	H	H	H	H	H		H	M	M	H	M	M	H	H	H	H	h	H	H	m	h	H	H	H	
sp						H	H	L	h	h	h	h	h	h	H			h				H			
sc							H	H	H	H	H	H		L		H	H	h	h	h		H	h		
gc							H	H	H	H	H	H	h	h	h	H	H					H			
hy	H	h					H	H	h																
ix							H	H	H	H	h	h	h	h	H	h		h	h		h	H	h		
ex							H	H	H	H	h	h	h	h	h				H			H			
ab							H	H	H	H	H	H	H	H	H	h			H	H		H		H	
mt	h	h	h	h	h	H	H	H	H	H	H	H	H	M	H	H	H	L	h	h	L	h			
ta							H	H	H	H												H			
ge							H	H	H	H	H	H	H	H	H	H	H		H	m	m	M	m	l	
cl	h	h	h	h	h	H	H	H	H	H	L	L	L	H	M	M	H	h	h			H			
qa						H	H	H	H	H	H	H	M	M	m	h	h	h	h	h		h	L	h	
su						H	H	H	H	H	H	M	M	m	h	h	h	H	h		H		H		

Applications label rows, and lexical information labels columns. Within a cell, the following indications appear:

- H heavy use of this feature
- M medium use
- L light use
- h heavy future use predicted
- l light future use

A blank cell indicates no current use of a lexical feature, and no surmisable future use (although the latter is clearly difficult to determine with any accuracy). On the next page an explanation of the abbreviations is given.

*Abbreviations for lexical information*

ph	phonetic transcription
sy	syllable indication
st	stress indication
du	duration indication
av	acoustic variant
or	orthography
mo	morpheme
ms	morphosyntax
ps	part of speech
sc	subcategorisation
ar	argument structure
th	theme
sr	semantic relation
sf	semantic feature
lr	lexical relation
co	collocation
id	idiom
tr	translation information
su	subject field
rh	rhetoric role
us	usage
pr	pragmatic
fr	frequency
di	discourse
dt	dialect

*Abbreviations for types of application*

sr	speech recognition
ss	speech synthesis
in	automatic interpretation
sp	spell checking
sc	style checking
gc	grammar checking
hy	automatic hyphenation
ix	automatic indexing
ex	automatic extracting
ab	automatic abstracting
mt	machine translation
ta	tagging
ge	generation
cl	CALL
qa	question-answering
su	text summarising

## References

- [1] Calzolari, N., "Computer-aided Lexicography: Dictionaries and Word Data Bases", in Batori, I.S., Lenders, W., Putschke, W. (eds.), *Computational Linguistics*, Walter de Gruyter, Berlin, 1989, 510-519.
- [2] Calzolari, N., Zampolli, A., "Lexical Databases and Textual Corpora: a Trend of Convergence between Computational Linguistics and Literary and Linguistic Computing", *Research in Humanities Computing*, Hockey, S., Ide, N. (eds.), Oxford University Press, Oxford, 1991, 273-307.
- [3] Formalisms Working Group, "WG Linguistic Formalisms Workplan", Internal Report EAG-FWG-WP, Saarbrücken, 1993.
- [4] Heid, U., McNaught, J. (eds.), "Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications", Eurotra-7 Final Report, Stuttgart, 1991.
- [5] Heylen, D. (ed.), "Collocations and the Lexicalisation of Semantic Information", Final Report ET-10/75, Utrecht, 1993.
- [6] Lexicon Working Group, "WG Computational Lexicon Workplan", Internal Report EAG-LWG-WP, Pisa, 1993.
- [7] McNaught, J., Caroli, F., Hellwig, P., "Possible Applications of Reusable Lexical Resources", Eurotra-7 Technical Report, Manchester, 1990.
- [8] Menon, B., Modiano, N., "Working Group on Computational Lexicons, Task 1 - Lexicon Architecture", Prefinal Technical Report EAG-LWG-T1.1, Paris, 1993.
- [9] Walker, D., Zampolli, A., Calzolari, N., *Automating the Lexicon: Research and Practice in a Multilingual Environment*, Proceedings of a Workshop held in Grosseto, Oxford University Press, Oxford, 1994.

# Outline of a Model for Lexical Databases\*

Nancy Ide<sup>†</sup>, Jacques Le Maitre<sup>‡</sup>, Jean Véronis<sup>†,‡</sup>

<sup>†</sup> Department of Computer Science, Vassar College

<sup>‡</sup> Groupe Représentation et Traitement des Connaissances,  
Centre National de la Recherche Scientifique

## Abstract

In this paper we show that previously applied data models are inadequate for lexical databases. In particular, we show that relational data models, including unnormalized models which allow the nesting of relations, cannot fully capture the structural properties of lexical information. We propose an alternative feature-based model which allows for a full representation of sense nesting and defines a factoring mechanism that enables the elimination of redundant information. We then demonstrate that feature structures map readily to an object-oriented data model and show how our model can be implemented in an object-oriented DBMS.

## 1 Introduction

For 25 years or so, computers have been used in the production of dictionaries. Initially, computers were used primarily for the first step in dictionary-making, that is, the gathering and accessing of examples of word use (see bibliography and surveys in Kipfer, 1983, and Landau, 1984). In the early 1980's, the *COBUILD* dictionary project took this use of computers to the extreme, by compiling a 20 million word corpus of English, generating extensive lists of examples from concordances of the corpus, and creating entries based almost exclusively on these examples (Sinclair, 1987). Computers have also been used extensively for word-processing and typesetting dictionary entries. However, in the last decade lexicographers and publishers began to explore more sophisticated computer involvement in the dictionary-making process, in particular, the use of databases of lexical information (including, for example, pronunciation, part of speech, definition, etymology, etc. for each word) for entry-making. The existence of a *lexical database* provides enormous potential: it can be used to update and maintain dictionaries and check coherency within a dictionary or across related dictionaries, as well as enable the exchange and sharing of information among projects and even the automatic generation of several variant printed versions of a

---

\* The work described in this paper has been carried out in the context of a joint project of the Department of Computer Science at Vassar College and the Groupe Représentation et Traitement des Connaissances of the Centre National de la Recherche Scientifique (CNRS), which is concerned with the construction and exploitation of a large lexical data base of English and French. An earlier version of this paper was presented at the conference "Intelligent Text and Image Handling" (RIA91) in Barcelona, Spain, April 1991.

<sup>†</sup> Department of Computer Science, Vassar College, Poughkeepsie, New York 12601, U.S.A.

<sup>‡</sup> Groupe Représentation et Traitement des Connaissances, Centre National de la Recherche Scientifique, 31, Chemin Joseph Aiguier, 13402 Marseille Cedex 09, France

dictionary (e.g., a full version, a concise version, and a pocket version) from a common data source. Beyond this, such databases can be used as the basis for electronic dictionaries on CD-ROM, and for on-line consultation by scholars.

Quite independently, during the 1970's and 80's computational linguists began to develop *computational lexicons* for natural language processing programs. Computational lexicons differ from dictionaries intended for human use in that they must contain much more explicit and specific linguistic information about words and phrases (for example, typical subjects and objects for verbs, semantic features for nouns such as inanimate, human, etc.), and must be encoded in strictly formal structures usable by computer programs. However, large computational lexicons are extremely difficult to develop from scratch, and in the early 1980's computational linguists found that the information typically contained in dictionaries could be exploited in many ways if it were organized in a computer database (see, for instance, Amsler, 1980; Calzolari, 1984; Markowitz, Ahlsweide & Evens, 1986; Byrd *et al.*, 1987; Nakamura & Nagao, 1988; Véronis & Ide, 1990; Klavans, Chodorow & Wacholder, 1990; Wilks *et al.*, 1990). No publishers' lexical databases existed at the time, but computational linguists saw that some relatively inexpensive processing of typesetter's tapes from printed dictionaries provided the means to create them. Consequently, lexical databases have been created from a number of printed dictionaries, including the *LDOCE*,<sup>1</sup> *Webster's 7th*, several Collins bilingual dictionaries, etc.

The goal of this paper is to provide a data model that is suited to lexical databases. Lexical data, as is obvious in any dictionary entry, is much more complex than the kind of data (suppliers and parts, employees' records, etc.) that has provided the impetus for most database research. Therefore, classical data models (e.g., relational) do not apply very well to lexical data, although several attempts have been made. In section 2, we review previously applied models and discuss their shortcomings, in an effort to better understand what is required to represent lexical data. A fundamental problem arises from the fact that dictionaries--and therefore databases--organize what seems to be the same kind of information (orthography, pronunciation, part of speech, etymology, definitions, etc.) in structurally different ways. A strong requirement for a data model is that it must make lexical information compatible despite this variability in structure. Compatibility is a necessary condition for data sharing and interchange, as well as for the development of general software.

In section 3, we show that a model based on *feature structures* overcomes most of the problems inherent in other models, and in particular enables accessing, manipulating, and merging information structured in multiple ways. The model

---

<sup>1</sup>In this paper we will use the following abbreviations for dictionary names:

<i>CED</i>	<i>Collins English Dictionary</i>
<i>LDOCE</i>	<i>Longman Dictionary of Contemporary English</i>
<i>NPEG</i>	<i>The New Penguin English Dictionary</i>
<i>OALD</i>	<i>Oxford Advanced Learner's Dictionary</i>
<i>OED</i>	<i>Oxford English Dictionary</i>
<i>SOED</i>	<i>Shorter Oxford English Dictionary</i>

allows retaining the particular organization of a given dictionary while at the same time making it transparent to certain database operations. There exists a well-developed theoretical framework for feature structures, which are widely used for information representation in linguistics. Their applicability to lexical databases seems therefore natural, although to our knowledge this has not yet been implemented. The use of feature structures in lexical databases also opens up the possibility of compatibility with computational lexicons, which is of considerable interest for computational linguists. A common representation scheme could create a useful bridge between lexicographers and computational linguists, and foster cross-fertilization between the two fields.

To date, feature-based models have not been exploited in commercial database management systems (DBMS), and therefore no implementation in such a system is possible. However, we show that feature structures map readily into object-oriented data models. In section 4 we describe an implementation of our model in the object-oriented DBMS O<sub>2</sub>.

## 2 Previous work

### 2.1 Text models

Dictionaries were first realized electronically as typesetters' tapes for the purposes of publishing. As mentioned above, these tapes subsequently became available to the research community, and several have been processed extensively in order to extract lexical information. In particular, the information about typographical rendering provided in typesetter's tapes (e.g., italics, bold, etc.) is replaced by labels which provide an indication of the content of a field (e.g., headword, pronunciation, etc.), rather than its printed form. In either case, a dictionary is seen as a strictly linear text stream interspersed with *markup* or *tags*, which we refer to as the *text model*. There have been some efforts to develop means for information retrieval from dictionaries in this form, since the fundamental textual nature of dictionaries does not immediately suggest the use of traditional database retrieval strategies.

#### 2.1.1 Typographical markup

Typographical markup signals a font shift or the presence of special characters, etc., corresponding to the rendering of the dictionary in printed form (Fig. 1). Markup of this kind is said to be *procedural*, because it specifies the procedure (e.g., shift to italic) to be performed when a tag is encountered in linear processing, rather than providing an indication of content of the field (see Coombs, Renear & DeRose, 1987). Although typographic codes are to some extent indicative of field content (for example, part of speech may always be in italics in a given dictionary), a straightforward, one-to-one mapping between typographic codes and content clearly does not exist, since other items, such as semantic field, usage, register, geographical information etc., may be rendered in italics as well. Positional

information can be coupled with typographic tags to determine content, but a complex analysis of entry format, which may or may not yield a definitive mapping due to ambiguities, is required. Information retrieval from a dictionary in this form is obviously costly, if it is possible at all.

```
*Cgin*E*S1*E (d*3T*3Fn) *Fn. *%brew *5Q*A1. *Ean alcoholic drink obtained by distillation and rectification of the grain of malted barley, rye, or maize, flavoured with juniper berries. *%brew *5Q*A2. *Eany of various grain spirits flavoured with other fruit or aromatic essences: *Fsloe gin. *%brew *5Q*A3. *Ean alcoholic drink made from any rectified spirit. *5Q*5HC18: shortened from Dutch *Fgenever *Ejuniper, via Old French form Latin *Fj=u-niperus*Gjuniper*E*5I<
```

Fig. 1. Typographical markup (*CED*)

### 2.1.2 Descriptive markup

In a *descriptive* markup scheme, tags provide an indication of the content of the fields they delimit rather than the printed rendering. For instance, instead of tags for italics, bold, etc., tags indicate *headword*, *part of speech*, *pronunciation*, etc. (Fig. 2). The use of descriptive markup enables the retrieval of information by content category from dictionaries in a linear text format. There have been a number of efforts to devise descriptive markup schemes for monolingual dictionaries (see, for example, Tompa, 1989) and to translate the procedural markup of typesetter's tapes into descriptive markup (see, for instance, Boguraev & Neff, in press). More recently, a preliminary common set of descriptive tags for encoding mono- and bi-lingual dictionaries has been proposed (Amsler & Tompa, 1988). This work has been incorporated into the international Text Encoding Initiative's guidelines for encoding machine readable textual and linguistic data (Sperberg-McQueen & Burnard, 1990).

```
<ent h=gin hn=2><hdw>gin</hdw><pr><ph>dZIn</ph></pr>
<hps ps=n cu=U><hsn><def>colourless alcoholic drink distilled
from grain or malt and flavoured with juniper berries, often
drunk with tonic water, and used in many kinds of
cocktail</def></hsn></hps></ent>
```

Fig. 2. Descriptive markup (*OALD*)

Sophisticated retrieval software for tagged text exists (for example, PAT; Gonnet, Baeza-Yates & Snider, 1991), which regards markup as strings of characters embedded in text and basically performs sophisticated string searches. However, texts manipulated by such software must be in a static form, and it is very costly to apply it to texts which are often modified or under development. Further, although such software provides powerful searching capabilities, it is nonetheless

limited for contextual searching. The software has no knowledge of the structure of the text, and so searches which involve elements whose relationship is embodied in the structure of the dictionary entry can become prohibitively complex. For example, part of speech is usually given once at the head of an entry, although it applies to all senses. Therefore, to find the part of speech for sense 4 of a given word, information that is physically distant from that sense in the text is required, which in turn demands analysis of the surrounding text. This is accomplished by specifying several string searches and applying Boolean operators to the results. Such operations are often complicated and unintuitive. Therefore, users must typically provide the right combination of requests by hand, and queries by external software are virtually prohibited.

### 2.1.3 Grammar-based models

The problems cited above have led to the development of a more sophisticated model, which superimposes a structure on the text stream by means of a context-free grammar describing the hierarchical structure of a document (Gonnet & Tompa, 1987). The text is represented in the form of "parsed strings", that is, the string itself together with its parse tree according to a given grammar (Fig. 3).

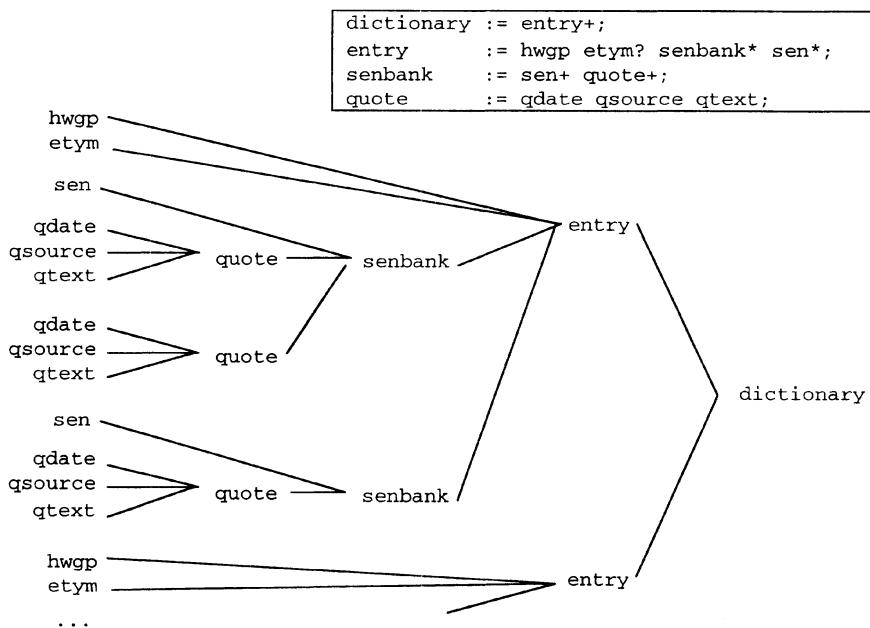


Fig. 3. Simplified grammar and a part of a "parsed-string" for the *OED* (from Gonnet & Tompa, 1987)

This model overcomes the contextual limitations of purely linear models, since the context is available in the parse tree. Also, Gonnet and Tompa (1987) show that it is possible to define the equivalent of many of the operators available in conventional databases with this model. However, although it has much potential interest, like linear models this model appears to be limited to fixed texts which cannot be easily modified or updated. Thus, the model may be applicable to publically distributed read-only dictionaries, but at this time seems unable to meet the requirements for lexical databases used in publishing and research.

## 2.2 Relational data models

Conventional data models for lexical databases have been proposed, primarily for the purposes of research in computational linguistics. These data models have been less popular with publishers and lexicographers, who have traditionally mistrusted such models as too simplistic and/or rigid to allow the editorial freedom lexicographers desire when creating dictionaries (Tompa, 1989).

At the present time, the most common data model is the *relational model*. A relational database consists of a set of *relations* between *entities*. Each role in that relation is called an *attribute*. Conceptually, a relation is a *table* whose columns correspond to *attributes*, and each row, or *tuple*, specifies all the values of attributes for a given entity. Attributes have only *atomic values*--that is, values which, from the DBMS's point of view, cannot be decomposed. In other words, each row-to-column intersection contains one, and only one, value.

The relational model has been suggested for representing dictionary information (see, for example, Nakamura & Nagao, 1988). In this scheme, the dictionary is represented by a set of relations, each of which includes attributes such as grammar codes, definitions, examples, etc. Fig. 4 gives the definition of "abandon" from the *LDOCE*; Fig. 5, expanded from Nakamura and Nagao (1988), shows the tabular representation of the same entry.<sup>2</sup>

**a·ban·don<sup>1</sup>** /@'bænd@n/ v [T1] 1 to leave completely and for ever; desert: *The sailors abandoned the burning ship.* 2 to leave (a relation or friend) in a thoughtless or cruel way: *He abandoned his wife and went away with all their money.* 3 to give up, esp. without finishing: *The search was abandoned when night came, even though the child had not been found.* 4 (to) to give (oneself) up completely to a feeling, desire, etc.: *He abandoned himself to grief | abandoned behaviour.* -- ~ment n [U].

**abandon<sup>2</sup>** n [U] the state when one's feelings and actions are uncontrolled; freedom from control: *The people were so excited that they jumped and shouted with abandon / in gay abandon.*

Fig. 4. Definition of 'abandon' from *LDOCE*

<sup>2</sup>Certain information, such as the *LDOCE* semantic "box codes", appears only in the machine-readable version of the dictionary, and it therefore appears in the database even though absent from the printed version.

## DEFINITION

HW	PS	DN	DF
abandon	v	1	to leave completely and for ever
abandon	v	1	desert
abandon	v	2	to leave (a relation or friend) in a thoughtless or cruel way
abandon	v	3	to give up, esp. without finishing
abandon	v	4	to give (oneself) up completely to a feeling, desire, etc.
abandon	n	0	the state when one's feelings and actions are uncontrolled
abandon	n	0	freedom from control

## EXAMPLE

HW	PS	DN	SP
abandon	v	1	The sailors abandoned the burning ship
abandon	v	2	He abandoned his wife and went away with all their money
abandon	v	3	The search was abandoned when night came, even though the child had not been found
abandon	v	4	He abandoned himself to grief
abandon	v	4	abandoned behaviour
abandon	n	0	The people were so excited that they jumped and shouted with abandon/in gay abandon

HW = headword

PS = part of speech

DN = definition number

DF = definition text

SP = example

GC = grammar code

BC = LDOCE "box" code

## CODE

HW	PS	DN	GC	BC
abandon	v	1	T1	----H---T
abandon	v	2	T1	--D-H---H
abandon	v	3	T1	----H---T
abandon	v	4	T1	----H---H
abandon	n	0	U	----T----

Fig. 5. Tables for 'abandon' in *LDOCE* database

This example is derived from a small, simple learner's dictionary with a straightforward internal structure (no deep nesting of senses, etc.), and several pieces of information from the entry text (for example, pronunciation, run-ons, cross-references) have been omitted from the database. However, even this simplified case shows that the relational model poses several problems for representing dictionary entries.

### 2.2.1 Fragmentation of data

The most obvious problem arises from the fact that the number of values for each attribute in dictionary entries varies enormously. For example, entries may include several different pronunciations, parts of speech, orthographic variants, definitions, etc., while some other fields, such as examples, synonyms, cross-references, domain information, geographical information, etc., may be completely absent. To avoid massive duplication of data, the information must be split across several tables, thus fragmenting the view of the data. The more complex the data, the more tables are required, and the more fragmented the view.

This fragmentation in a relational database is real--that is, the different relations represented in different tables are not explicitly connected, but are only logically

connected by including attributes with the same domains in the different tables (for example, HW, PS, and DN in Fig. 5). Tuples are connected only if the values for those attributes match. As a result, queries can become complex and unintuitive. For example, Fig. 6 shows the SQL query required to extract all examples given for uncountable nouns whose definitions start with the string "the state...", as for the noun sense of *abandon*.

```

SELECT DEFINITION.HW, EXAMPLE.SP
FROM DEFINITION, EXAMPLE, CODE
WHERE
    DEFINITION.PS = "n" AND
    DEFINITION.DF = "the state *" AND
    EXAMPLE.SP <> null AND
    CODE.GC = "U" AND
    DEFINITION.HW = EXAMPLE.HW AND
    EXAMPLE.HW = CODE.HW AND
    DEFINITION.PS = EXAMPLE.PS AND
    EXAMPLE.PS = CODE.PS AND
    DEFINITION.DN = EXAMPLE.DN AND
    EXAMPLE.DN = CODE.DN

```

Fig. 6. Sample SQL query

HW	PS	DN	GC	BC	DF	SP
abandon	v	1	T1	---H---T	to leave completely and for ever	The sailors abandoned the burning ship
abandon	v	1	T1	---H---T	desert	The sailors abandoned the burning ship
abandon	v	2	T1	--D-H--H	to leave (a relation or friend) in a thought- less or cruel way	He abandoned his wife and went away with all their money
abandon	v	3	T1	---H---T	to give up, esp. without finishing	The search was abandoned when night came, even though the child had not been found
abandon	v	4	T1	---H---T	to give (oneself) up completely to a feeling, desire, etc.	He abandoned himself to grief
abandon	v	4	T1	---H---T	to give (oneself) up completely to a feeling, desire, etc.	abandoned behaviour
abandon	n	0	U	---T---	the state when one's feelings and actions are uncontrolled	The people were so excited that they jumped and shouted with abandon/in gay abandon
abandon	n	0	U	---T---	freedom from control	The people were so excited that they jumped and shouted with abandon/in gay abandon

Fig. 7. Joined table from DEFINITION, EXAMPLE and CODE in Fig. 5.

A less fragmented view can be obtained by *joining* two or more tables, but the resulting table typically contains an enormous amount of redundant information. Fig. 7 shows the result of joining the tables from Fig. 5, and although very little of the information in an actual dictionary entry is represented here, it is already

obvious that the resulting view is cumbersome. Of course, joining makes query construction easier, but the burden is then on the user to appropriately define the most commonly used views, and on the management system to handle the processing.

### 2.2.2 Hierarchical structure

Whether a view is fragmented or joined, there is no representation in a relational database of the obvious hierarchy within most dictionary entries—for instance, it is clear that the entry for *abandon* has two main sub-parts, one for its verb senses and one for its noun sense, and that the two senses of the verb labeled "1" in Fig. 5 are in fact two sub-senses of the first sense given in the entry. These two sub-senses are more closely related to each other than to senses 2, 3, and 4, but the tabular format obscures this fact. Some dictionaries take the grouping and nesting of senses several levels deep in order to distinguish finer and finer grains of meaning. The Hachette Zyzomys CD-ROM dictionary, for instance, distinguishes up to five levels in an entry. Fig. 8 shows that in this dictionary "valeur" has two fundamental senses: (A) *value* as *merit*; and (B) *value* as price. Going deeper, we see that sense A subdivides into two main subcategories: (I) merit of an individual; and (II) the subjective worth of an object. Sense A.I subdivides further into two more subcategories: (1) merit of a person based on general qualities; and (2) bravery or valor, which in turn forms a part of the compound "*croix de la valeur militaire*", a French military decoration. Flattening this structure into a tabular form obscures the derivational relations captured in the nested arrangement.

**valeur** [valœR] n. f. A. I. 1. Ce par quoi une personne est digne d'estime, ensemble des qualités qui la recommandent. (V. *mérite*). *Avoir conscience de sa valeur. C'est un homme de grande valeur.* 2. Vx. Vaillance, bravoure (spécial., au combat). "La valeur n'attend pas le nombre des années" (Corneille). ♦ *Valeur militaire (croix de la)*: décoration française...  
...  
**II. 1.** Ce en quoi une chose est digne d'intérêt. *Les souvenirs attachés à cet objet font pour moi sa valeur.* **2.** Caractère de ce qui est reconnu digne d'intérêt...  
...  
**B. I. 1.** Caractère mesurable d'un objet, en tant qu'il est susceptible d'être échangé, désiré, vendu, etc. (V. *prix*). *Faire estimer la valeur d'un objet d'art...*

Fig. 8. Part of the definition of 'valeur' in Hachette Zyzomys

### 2.3 Unnormalized relational models

The need to eliminate redundancy by factoring common pieces of information is well known in database research and has led to the development of *unnormalized* (also *Non First Normal Form* or *NF<sup>2</sup>*) relational data models, in which attribute values may have a composite structure. That is, attribute values may be either atomic, as in the classical *normalized* relational model, or they may be nested relations with their own internal structure. An algebra and calculus have been

proposed for these relations (Abiteboul & Bidoit, 1984; Roth, Korth & Silberschatz, 1988), and a few DBMSs have been developed using the NF<sup>2</sup> model (e.g., VERSO [Bancilhon, 1983], AIM-P [Pistor & Traunmueller, 1986], DASDBS [Schek, Paul, Scholl & Weikum, 1990]).

Fig. 9 shows how the entry for *abandon* given in Fig. 4 would be represented in the NF<sup>2</sup> model. The outermost table consists of a relation between a headword and some number of homographs. In turn, a homograph consists of a part of speech, a grammar code, and some number of senses; etc. Obviously, this model better captures the hierarchical structure of information in the dictionary and enables the factoring of attributes. Relations which were represented only in matched attribute values between tables in the normalized model are now made explicit, and therefore, complex queries do not require specification of the table connections.

HW	HOMOGRAPH							
	PS	GC	SENSE					
			DN	BC	DEFINITION		EXAMPLE	
					DF		SP	
abandon	v	T1	1	----H---T	to leave completely and for ever <i>desert</i>		The sailors abandoned the burning ship	
			2	--D-H---H	to leave (a relation or friend) in a thought- less or cruel way		He abandoned his wife and went away with all their money	
			3	----H---T	to give up, esp. without finishing		The search was abandoned when night came, even though the child had not been found	
			4	----H---T	to give (oneself) up completely to a feeling, desire, etc.		He abandoned himself to grief <i>abandoned behaviour</i>	
			n	U	0	----T----	the state when one's feelings and actions are uncontrolled <i>freedom from control</i>	The people were so excited that they jumped and shouted with abandon/in gay abandon

Fig. 9. NF<sup>2</sup> representation of the entry 'abandon'

Neff, Byrd, and Rizk (1988) describe an organization for a lexical database (the IBM LDB) (see also Calzolari, Peters & Roventini, 1990) also based on a hierarchy of attributes, which allows the representation of information in a dictionary entry as a tree (see Fig. 10). Queries are made by filling templates with the same tree structure as the dictionary entry and by indicating desired values and conditions for various attributes (Fig. 11). Therefore, the IBM LDB model is fundamentally an NF<sup>2</sup> model, although it does not use an NF<sup>2</sup> DBMS but is instead built around an *ad hoc* implementation.

Although NF<sup>2</sup> models clearly improve on other models for representing dictionary information, a number of problems still remain. These are outlined in the following sub-sections.

```

entry
+-hdw: abandon
|
+-superhom
| +-word: abandon
| +-print_form: a.ban.don
| +-hom_number: 01
|
| +-pronunciation
| | +-primary
| |   +-pron_string: E"b&ndEn
|
| +-syncatid: v
| +-g_code_field: T1
|
| +-sense_def
| | +-sense_no: 1
| | +-subj_code: ....
| | +-box_code: ....H....T
|
| +-defn
| | +-def_string: to leave completely and for ever; desert
|
| +-example
|   +-ex_string: The sailors abandoned the burning ship
|
| +-sense_def
| | +-sense_no: 2
...
...
| +-run_on
|   +-sense_link: 01
|   +-derivative: abandonment
|   +-d_syncatid: n
|
| +-d_code
|   +-g_code_field: U
|
+-superhom
  +-word: abandon
  +-print_form: abandon
  +-hom_number: 02
  +-syncatid: n
  +-g_code_field: U
|
+-sense_def
  +-sense_no: 0
  +-subj_code: ....
  +-box_code: ....T.....
|
| +-defn
|   +-def_string: the state when one's feelings and actions...
|
| +-example
|   +-ex_string: The people were so excited that they jumped...

```

Fig. 10. IBM LDB format for 'abandon' in the *LDOCE*

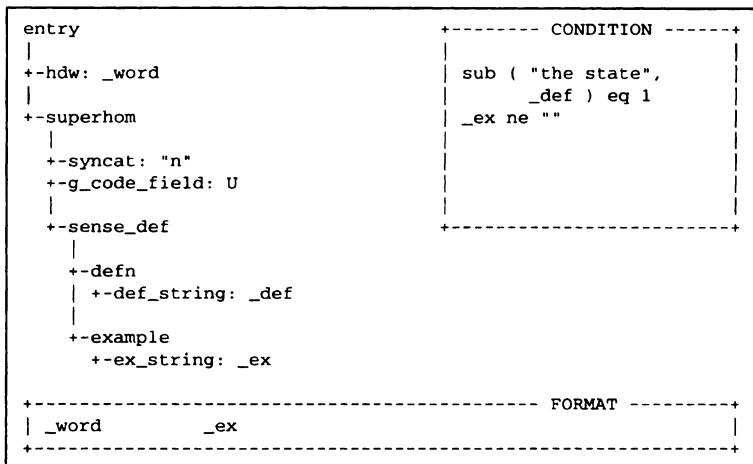


Fig. 11. IBM LDB -- sample query

### 2.3.1 Recursive nesting

Fig. 12 shows an attempt to render the entry for *valeur* from the *Zyzomys* dictionary given in Fig. 8 in the IBM LDB format. It is clear that the IBM LDB model and NF<sup>2</sup> models in general can represent the deep hierarchical structure of the entry. However, NF<sup>2</sup> models do not allow the recursive nesting of relations, and Neff, Byrd, and Rizk (1988) explicitly prohibit recursion in the IBM LDB model. This necessitates the proliferation of attributes such as *sense\_def\_level\_1*, *sense\_def\_level\_2*, etc. to represent the different levels of sense nesting. This in turn demands that queries take into account all the possible positions where a given sub-attribute (e.g., *usage*) could appear. For example, all the queries in Fig. 13 are required to retrieve all nouns which have an archaic (*Vx = vieux*) sense. Since any sense at any level could have this attribute value, it is necessary to query each level.

### 2.3.2 Exceptions

Exceptional cases are characteristic of lexical data. For instance (see Fig. 14):

- Sense 3 of the word "conjure" in the *OALD*, has a different pronunciation from the other senses in the entry.
- In the same entry, the related entry "conjuror, conjuror", although given at the entry level, applies only to sense 1.<sup>3</sup>

---

<sup>3</sup>Note that the dictionary has to rely on a special mechanism ("1 above") to make this specification.

```

entry
+-hdw: valeur
|
+-superhom
| +-word: valeur
...
|
| +-sense_def_level_1:
|   +-sense_no: A
|   +-sense_def_level_2:
|     +-sense_no: I
|     +-sense_def_level_3:
|       +-sense_no: 1
|       +-defn
|         +-def_string: Ce par quoi une personne est digne d'estime...
|
|       +-example
|         +-ex_string: Avoir conscience de sa valeur.
|         +-ex_string: C'est un homme de grande valeur.
|
|   +-sense_def_level_3:
|     +-sense_no: 2
|     +-usage: Vx
|     +-defn
|       +-def_string: Vaillance, bravoure (spécial., au combat)
|     +-example
|       +-ex_string: "La valeur n'attend pas le nombre des années"
|       +-ex_author: Corneille
|
|   +-sense_def_level_4:
|     +-sense_no: a
|     +-compound: Valeur militaire (croix de la)
|     +-defn
|       +-def_string: décoration française...
...

```

Fig. 12. Attempt to render the entry *valeur* in the IBM LDB format

<pre> entry   +-hdw: _word   +-superhom     +-syncat: "n"   +-sense_def_level_1:   +-usage: "Vx" </pre>	<pre> entry   +-hdw: _word   +-superhom     +-syncat: "n"   +-sense_def_level_1:   +-sense_def_level_2:     +-usage: "Vx" </pre>	<pre> entry   +-hdw: _word   +-superhom     +-syncat: n   +-sense_def_level_1:   +-sense_def_level_2:     +-sense_def_level_3:       +-usage: Vx </pre>
---	--	---

etc.

Fig. 13. Query problem

- The entry "heave" in both the *OALD* and *CED* shows that inflected forms may apply to individual senses--in this case, the past tense and past participle is "heaved" for all but the nautical senses, for which it is "hove".<sup>4</sup>
- The entry for "breath" from the *SOED* and the entry for "silk" from *NPEG* each specify a special etymology for a particular sense within their respective entries.

**con•jure** /'k<sup>nd</sup>G@(r)/ *vt, vi* 1 [VP2A,15A] do clever tricks which appear magical...  
**2** [VP15B] ~ *up*, cause to appear as if from nothing... 3 /k@n'dGW@(r)/ [VP17] (formal) appeal solemnly to... **con•juror, con•juror** /'k<sup>nd</sup>G@r@(r)/ *n* person who performs conjuring tricks } 1 above. [OALD]

**heave** /hi:v/ *vt, vi* (*pt, pp ~d* or (6 and 7 below), nautical use, *hove* /h@Wv/) ... [OALD]

**heave** (hi:v) *vb. heaves, heaving, heaved or (chiefly nautical) hove.* ... 5. (*past tense and past participle hove*) *Nautical.* a. to move or cause to move in a specified way ... b. (*intr.*) (of a vessel) to pitch or roll. ... [CED]

**Breath** (breP). [OE. *bræP* odour, exhalation ... The sense 'air in the lungs or mouth' was taken over from OE. *æPm* and *anda* (ME *ethem* and *onde*).] ... [SOED]

**silk** /silk/ *n* 1 a fine continuous protein fibre ... 3 a King's or Queen's counsel ... [ME, fr OE *seolc* ... (3) fr the silk gown worn by a King's or Queen's counsel] [NPED]

Fig. 14. Exceptions in dictionary entries

Allowing the same attribute at different levels, in different nested relations (for example, allowing a pronunciation attribute at *both* the homograph and sense levels) would require a mechanism to "override" an attribute value at an inner level of nesting. NF<sup>2</sup> models do not provide any such mechanism and, in fact, do not allow the same attribute to appear at different levels. The only way exceptions could be handled in an NF<sup>2</sup> model would be by re-defining the template so that attributes such as pronunciation, inflected forms, etymology, etc., are associated with senses *rather than* homographs. However, this would disable the factoring of this information, which applies to the entire entry in the vast majority of cases. The result would be an effective return to the normalized model.

### 2.3.3 Variable factoring

Dictionaries obviously differ considerably in their physical layout. For example, in one dictionary, all senses of a given orthographic form with the same etymology

---

<sup>4</sup>Again, a special rendering mechanism is required to handle this case, since it so grossly violates the usual entry format.

will be grouped in a single entry, regardless of part of speech; whereas in another, different entries for the same orthographic form are given if the part of speech is different. The *CED*, for instance, has only one entry for *abandon*, including both the noun and verb forms, but the *LDOCE* gives two entries for *abandon*, one for each part of speech. As a result of these differences, the IBM LDB template for the *LDOCE* places the part of speech (*syncat*) attribute at the homograph level, whereas in the *CED* template, part of speech must be given at the level of sense (or "sense group" if some new attribute were defined to group senses with the same part of speech within an entry). This means that the query for part of speech in the *LDOCE* is completely different from that for the *CED*. Further, it means that the merging or comparison of information from different dictionaries demands complete (and possibly complex) de-structuring and re-structuring of the data. This makes data sharing and interchange, as well as the development of general software for the manipulation of lexical data, difficult.

However, differences in dictionary layout are mainly differences in structural organization, whereas the fundamental elements of lexical information seem to be constant. In the example above, for instance, the basic information (orthography, pronunciation, part of speech, etc.) is the same in both the *CED* and *LDOCE*, even if its organization is different. Recognizing this, there have been various efforts to develop a taxonomy of lexical data and a meta-lexical terminology that can generalize across dictionaries and projects (Brustkern & Hess, 1982; The DANLEX Group, 1987).

The only way to have directly compatible databases for different dictionaries in the  $NF^2$  model, even if one assumes that attributes for the same kind of information (e.g., orthography) can have the same *name* across databases, is to have a common *template* across all of them. However, the fixed factoring of attributes in  $NF^2$  models prohibits the creation of a common template, because the template for a given database mirrors the particular factoring of a single dictionary. Therefore, a more flexible model is needed that would retain the particular factoring of a given dictionary, and at the same time render that factoring transparent to certain database operations.

### 3 A feature-based model

In this section we introduce a model for representing information in dictionary entries based on *feature structures*. Feature structures have been heavily used in formal and computational linguistics and natural language processing to encode linguistic information, especially in various grammar formalisms (see, for instance, Kaplan & Bresnan, 1982; Kay, 1985). Their applicability to the information found in dictionaries seems natural and opens up the possibility of compatibility with computational lexicons, although to our knowledge feature structures have not yet been used to represent dictionaries. In addition, there exists a well-developed theoretical framework for the feature structure mechanism which can provide a basis for the model we develop here.

### 3.1 Feature structures

In this section, we give a very brief overview of feature structures. For a more detailed introduction we refer the reader to Shieber (1986). A feature structure is composed of pairs of attributes (called *features*) and their values, which can also be seen as *partial functions* from features to values. Feature structures are graphically represented as a list of features separated from their values by colons, enclosed in square brackets (see Fig. 15a). Values may be atomic or may themselves be feature structures (Fig. 15b and c).

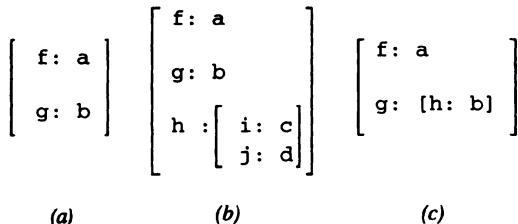


Fig. 15. Feature structure notation

A partial order relation called *subsumption* is defined for feature structures, which determines whether one feature structure is more general than another. A feature structure A is said to *subsume* another feature structure B (noted  $A \sqsubseteq B$ ) if for each feature  $f$  of A, there is a feature  $f$  in B and if the two values  $A(f)$  and  $B(f)$  are atomic then  $A(f) = B(f)$ , otherwise  $A(f) \sqsubseteq B(f)$ . Thus the feature structure (a) subsumes the feature structure (b) in Fig. 15.

An operation called *unification* is also defined to enable the combination of information from feature structures having different, but compatible, information. In formal terms, the unification of A and B (noted  $A \sqcup B$ ) is the greatest lower bound of A and B according to the subsumption relation--that is, the most general feature structure that is subsumed by both A and B. The dual operation, which consists of taking the least upper bound--that is, the most precise feature structure that subsumes both A and B--is called *generalization* (noted  $A \sqcap B$ ) (Fig. 16).

Two feature structures are said to be *compatible* if they can be unified, that is, if there exists a feature structure subsumed by both A and B. Otherwise, they are said to be *incompatible*. The feature structure (c) in Fig. 15 is incompatible with both (a) and (b).

A model based on feature structures can be used to represent simple dictionary entries, as shown in Fig. 17. Our model is *typed* in the sense that not all features can appear anywhere, but instead, must follow a schema that specifies which features are *allowable* (although not necessarily present), and where (see, for instance, Pollard & Sag, 1987). The schema also specifies the domain of values, atomic or complex, allowed for each of these features. For example, entries are described by the type ENTRY, in which the features allowed are *form*, *gram*, *usage*, *def*, etc. The domain of values for *form* is feature structures of type FORM, which

consists of feature structures whose legal features include *orth*, *hyph*, and *pron*. Each of these features has, in turn, an atomic value of type STRING, etc.

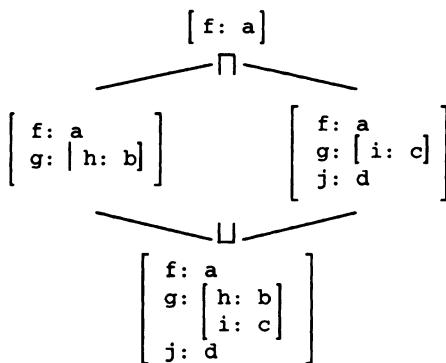


Fig. 16. Unification and generalization

**com•peti•tor/k@m'petɪtə(r)/ n person who competes [OALD]**

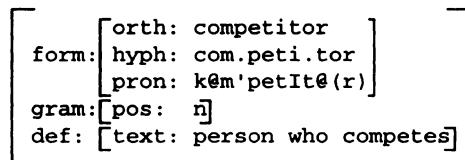


Fig. 17. Representation of a simple sense

### 3.2 Value disjunction and variants

The basic feature-based formalism is not enough to represent the structure of more complex dictionary entries. Fortunately, several authors have proposed extensions that solve many of the remaining problems. Karttunen (1984) shows that *value disjunction* is a natural, linguistically motivated extension, which enables the specification of a set of alternative values, atomic or complex, for a given feature. The use of value disjunction enables the representation of variants, common in dictionary entries, as shown in Fig. 18. We have added a further extension which enables the specification of either a set (noted  $\{x_1, \dots, x_n\}$ ) or a list (noted  $(x_1, \dots, x_n)$ ) of possible values. This extension enables retaining the order of values, which is in many cases important in dictionaries. For example, the orthographic form given

first is most likely the most common or preferred form. Other information, such as grammatical codes, may not be ordered<sup>5</sup>.

**biryani or biriani** (bɪr'ɪA:nɪ) *n.* Any of a variety of Indian dishes... [CED]

form:	[orth: (biryani, biriani)]
pron:	,biri'ɪA:nɪ
gram:	[pos: n]
def:	[text: Any of a variety of Indian dishes...]

Fig. 18. Value disjunction

In many cases, sets or lists of alternatives are not single values but instead *groups* of features. This is common in dictionaries; for instance, Fig. 19 shows a typical example where the alternatives are *groups* consisting of orthography and pronunciation.

**mackle** ('mæk@l) *or* **macule** ('mækju:l) *n.* *Printing.* a double or blurred impression caused by shifting paper or type. [CED]

form:	[{orth: mackle pron: 'm&k@l}, {orth: macule pron: 'm&kju:l}]
gram:	[pos: n]
usage:	[dom: Printing]
def:	[text: a double or blurred impression...]

Fig. 19. Value disjunction of non-atomic values

### 3.3 General disjunction and factoring

Kay (1985) proposes an additional extension, called *general disjunction*, that provides a means to specify alternative *sub-parts* of a feature structure. Again, we have extended the mechanism to enable the specification of both sets and lists of sub-parts. Therefore, feature structures can be described as being of the form  $[\phi_1, \dots, \phi_n]$ , where each  $\phi_i$  is a feature-value pair  $f: \psi$ , a set of feature structures  $\{\psi_1, \dots, \psi_p\}$ , or a list of feature structures  $(\psi_1, \dots, \psi_p)$ . Unification can be extended to disjunctive feature structures: if  $F = \{\phi_1, \dots, \phi_n\}$  and  $G = \{\psi_1, \dots, \psi_p\}$ , computing

---

<sup>5</sup>Since all of our examples are drawn from existing dictionaries, we have chosen to retain the original ordering rather than make decisions concerning the relevance of the order in which items appear. Therefore, mainly lists appear in the examples given here.

$F \sqcup G$  involves taking the disjunction of all the  $\phi_i \sqcup \psi_j$ , and then discarding all the nonmaximal disjuncts.

General disjunction allows common parts of components to be factored. Fig. 20a shows that without any disjunction, two different representations for the entry for *hospitaller* from the CED are required. The use of value disjunction enables localizing the problem and thus eliminates some of the redundancy (Fig. 20b), but only general disjunction (Fig. 20c) captures the obvious factoring and represents the entry cleanly and without redundancy.

**hospitaller or U.S. hospitaler** (*hospIt@l@*) *n.* a person, esp. a member of certain religious orders... [CED]

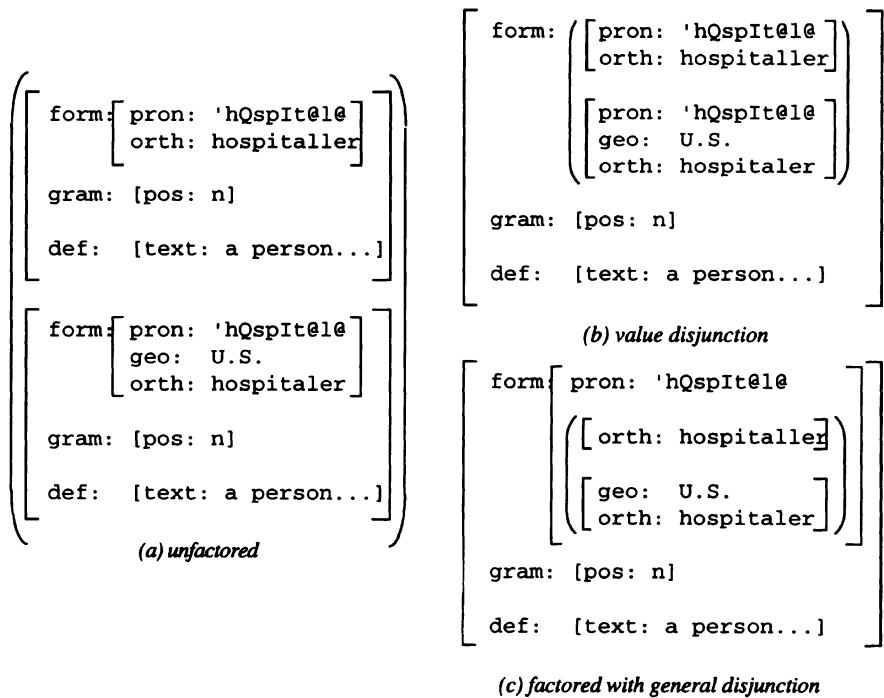


Fig. 20. General disjunction

General disjunction provides a means to represent multiple senses, since they can be seen as alternatives (Fig. 21).<sup>6</sup>

<sup>6</sup>Note that in our examples, "/" signals the beginning of a comment which is not part of the feature structure. We have not included the sense number as a feature in our examples because sense numbers can be automatically generated.

**disproof** (dis'pru:f) *n.* 1. facts that disprove something. 2. the act of disproving. [CED]

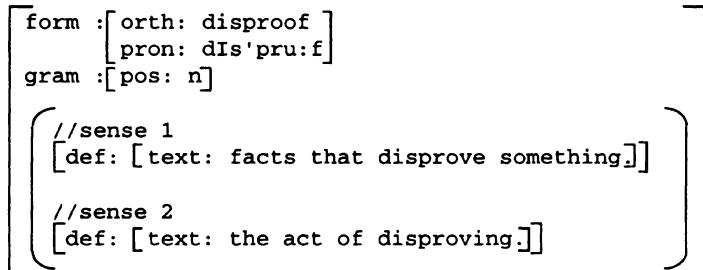


Fig. 21. Representation of multiple senses

Sense nesting is also easily represented using this mechanism. Fig. 22 shows the representation for *abandon* given previously. At the outermost level of the feature structure, there is a disjunction between the two different parts of speech (which appear in two separate entries in the *LDOCE*). The disjunction enables the factoring of orthography, pronunciation, and hyphenation over both homographs.<sup>7</sup> Within the first component of the disjunction, the different senses for the verb comprise an embedded list of disjuncts. Note that in this model there is no different type of feature structure for entries, homographs, or senses, since they potentially contain the same kind of information, as the discussion in section 2.3 demonstrates. This reflects a fundamental property of lexical data which is obscured by the layout of print dictionaries.

The entry for *valeur* from the *Zyzomys* dictionary provides an even more complex example of sense nesting (Fig. 23).

Note that we restrict the form of feature structures in our model to a *hierarchical normal form*. That is, in any feature structure  $F = [\phi_1, \dots, \phi_n]$ , only one  $\phi_i$ , let us say  $\phi_n = \{\psi_1, \dots, \psi_p\}$ , is a disjunction. This restriction is applied recursively to embedded feature structures. This scheme enables representing a feature structure as a tree in which factored information  $[\phi_1, \dots, \phi_{n-1}]$  at a given level is associated with a node, and branches from that node correspond to the disjuncts  $\psi_1, \dots, \psi_p$ . Information associated with a node applies to the whole subtree rooted at that node. For example, the tree in Fig. 24 represents the feature

<sup>7</sup> Interestingly, the *LDOCE* gives a separate entry for each part of speech, but gives information about hyphenation and pronunciation only in the entry for the first homograph. This shows that entries are an artifact of printed presentation and do not entirely reflect logical structure. The IBM LDB representation of the *LDOCE* entry for *abandon* loses the information about hyphenation and pronunciation for the second homograph, since there is no provision for the factoring of information in this scheme. The only solution in that model would be to repeat the information for the second homograph.

structure for *abandon* given in Fig. 22. The representation of information as a tree mirrors the hierarchical structure and factoring of information in dictionaries.

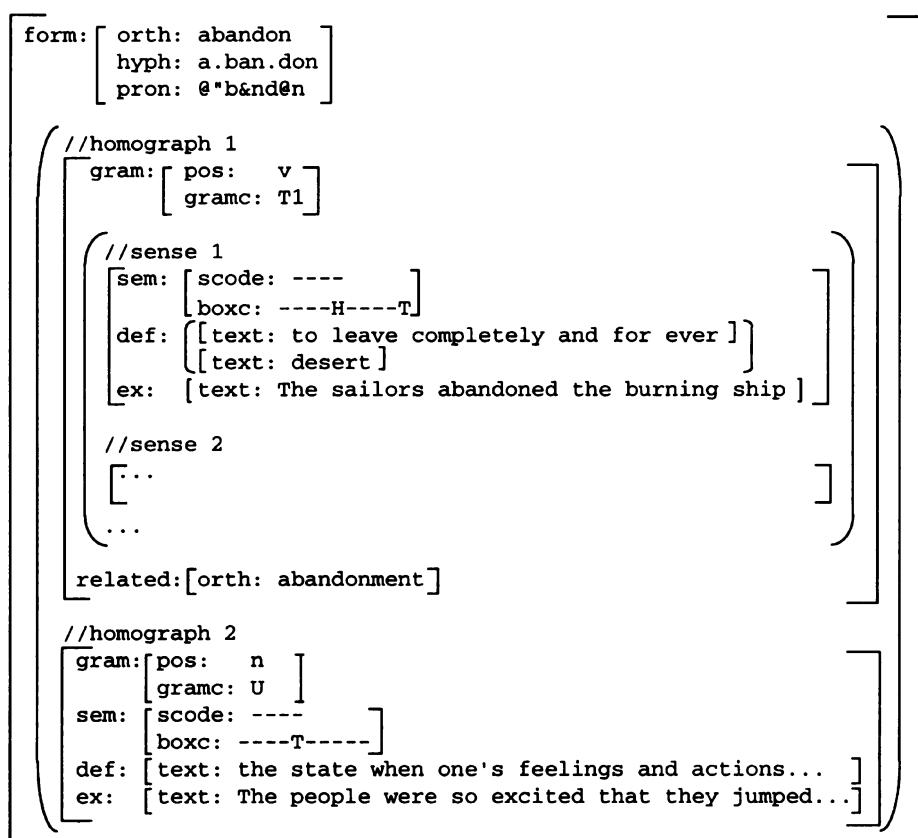


Fig. 22. Representation of the entry *abandon* in *LDOCE*

```

form: [orth: valeur]
      [pron: valeR]
gram: [pos: n]
      [gend: f]

  //sense A
  //sense A.I
    //sense A.I.1
      def: [text: Ce par quoi une personne est digne d'estime...]
      xref: [orth: mérite]
      ex:  [[text: Avoir conscience de sa valeur. ]
            [text: C'est un homme de grande valeur.]]
    //sense A.I.2
      time: Vx
      def: [text: Vaillance, bravoure (spécial., au combat).]
      ex:  [[text: La valeur n'attend pas le nombre des années]
            [auth: Corneille]
            [related: [orth: croix de la valeur militaire]]]
    ...
  //sense A.II
    //sense A.II.1
      def: [text: Ce en quoi une chose est digne d'intérêt.]
      ex:  [text: Les souvenirs attachés à cet objet...]
    //sense A.II.2
      def: [text: Caractère de ce qui est reconnu digne...]
    ...
  //sense B
  //sense B.I
    //sense B.I.1
      def: [text: Caractère mesurable d'un objet...]
      xref: [orth: prix]
      ex:  [text: Faire estimer la valeur d'un objet d'art.]
    ...

```

Fig. 23. Representation of the entry *valeur* in Zyzomys

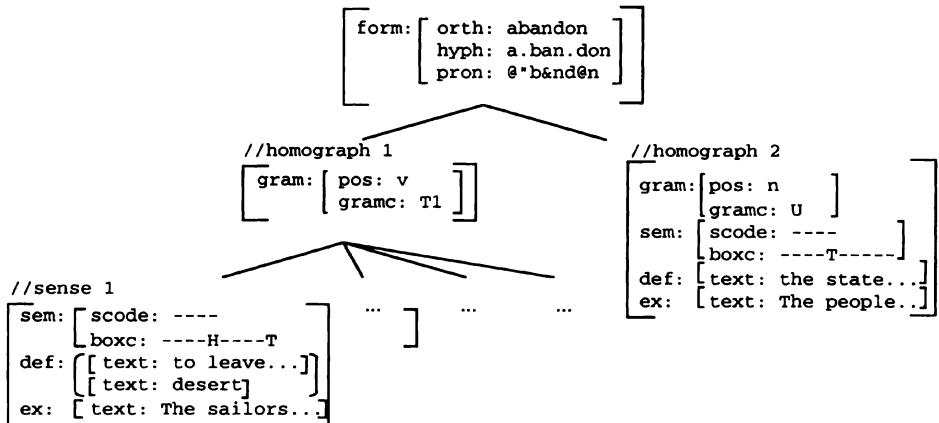


Fig. 24. Hierarchical Normal Form

### 3.4 Disjunctive normal form and equivalence

It is possible to define an *unfactor* operator to multiply out the terms of alternatives in a general disjunction (Fig. 25), assuming that no feature appears at both a higher level and inside a disjunct.<sup>8</sup>

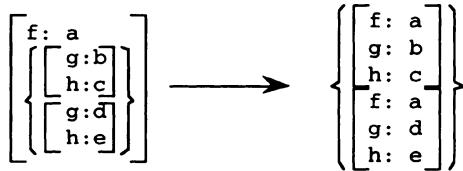


Fig. 25. Unfactoring

By applying the unfactor operator recursively, it is possible to eliminate all disjunctions except at the top level. The resulting (extremely redundant) structure is called the *disjunctive normal form* (DNF). We say that two feature structures are *DNF-equivalent* if they have the same DNF. The fact that the same DNF may have two or more equivalent factorings enables the representation of different factorings in dictionaries, while retaining a means to recognize their equivalence. Fig. 26a shows the factoring for inflected forms of *alumnus* in the CED; the same information could have been factored as it appears in Fig. 26b. Note that we have used *sets* and not *lists* in Fig. 26. Strictly speaking, the corresponding feature structures with lists would not have the same DNFs. However, since it is trivial to convert lists into sets, it is easy to define a stronger version of DNF-equivalence that disregards order.

<sup>8</sup>Value disjunction is not affected by the unfactor process. However, a value disjunction [f: {a, b}] can be converted to a general disjunction [{[f: a], [f: b]}], and subsequently unfactored.

**alumnus** (@'l^mn@s) or (fem.) **alumna** (@'l^mn@) n., pl. **-ni** (-nai) or **-nae** (-ni:) ... [CED]

(a)

**alumnus** (@'l^mn@s), pl. **-ni** (-nai), or (fem.) **alumna** (@'l^mn@), pl. **-nae** (-ni:)

(b)

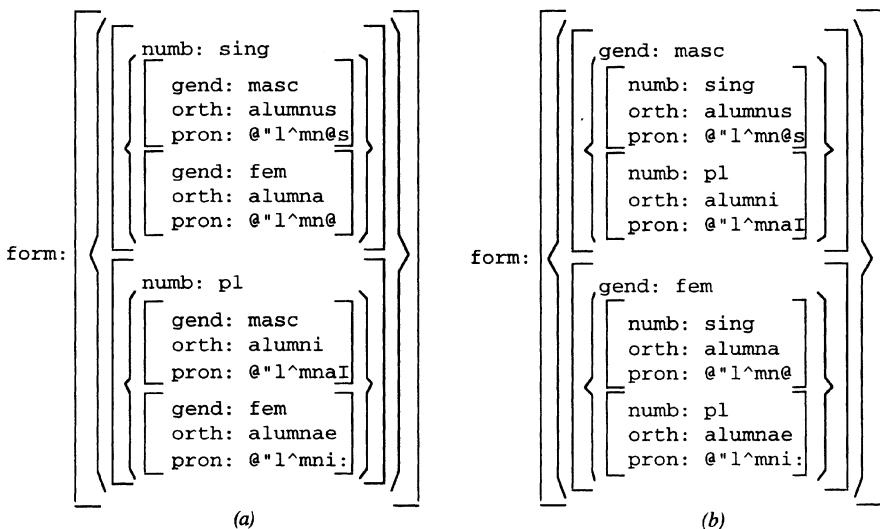


Fig. 26. Two different factorings of the same information

We can also define a *factor* operator to apply to a group of disjuncts, in order to factor out common information. Information can be unfactored and re-factored in a different format without loss of information, thus enabling various presentations of the same information, which may, in turn, correspond to different printed renderings or "views" of the data.

### 3.5 Partial factoring

The type of factoring described above does not handle the example in Fig. 27, where only a part of the grammatical information is factored (*pos* and *subc*, but not *gcode*). We can allow a given feature to appear at both the factored level and inside the disjunct, as long as the two values for that feature are *compatible*. In that case, unfactoring involves taking the *unification* of the factored information and the information in the disjunct (Fig. 28).

**ca•reen** /k@'ri:n/ *vt, vi* 1 [VP6A] turn (a ship) on one side for cleaning, repairing, etc. 2 [VP6A, 2A] (cause to) tilt, lean over to one side. [OALD]

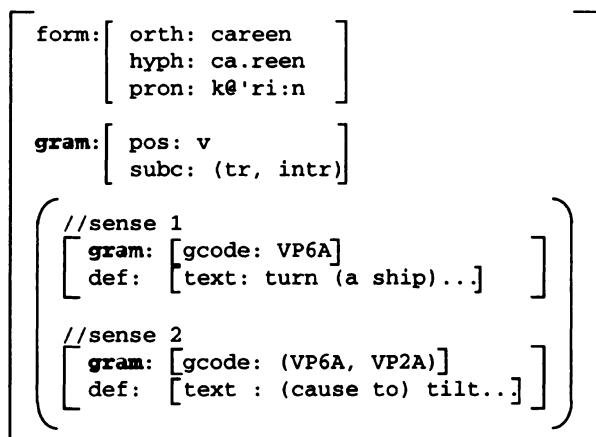


Fig. 27. Partial factoring

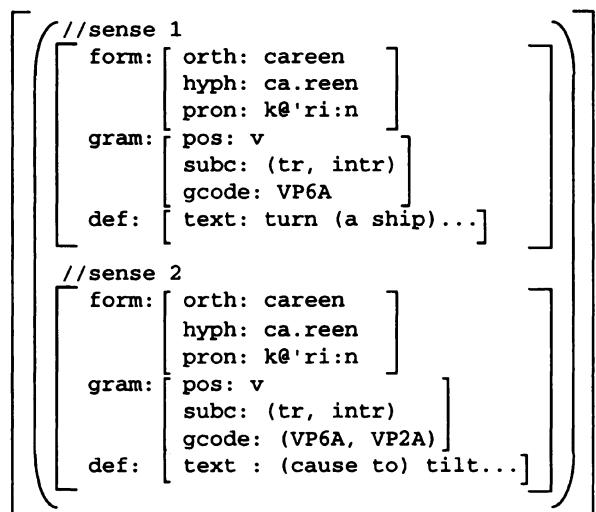


Fig. 28. Unfactored version of the feature structure in Fig. 27

### 3.6 Exceptions and overriding

We saw in the previous section that compatible information can appear at various levels in a disjunction. Exceptions in dictionaries will be handled by allowing *incompatible* information to appear at different levels. When this is the case,

unfactoring will be defined to *retain only the information at the innermost level*. In this way, a value specified at the outer level is *overridden* by a value specified for the same feature at an inner level. For example, Fig. 29 shows the factored entry for *conjure*, in which the pronunciation specified at the outermost level applies to all senses except sense 3, where it is overridden. Fig. 30 gives the unfactored version of the entry.

**con•jure** /'k<sup>ə</sup>ndʒə(r)/ *vt, vi* 1 [VP2A,15A] do clever tricks which appear magical...  
 2 [VP15B] ~ *up*, cause to appear as if from nothing... 3 /k@n'dʒWə(r)/ [VP17]  
 (formal) appeal solemnly to... [OALD]

```

form: [ orth: conjure
        hyph: con.jure
        pron: "kVndz@(r)"]
gram: [ pos: v
        subc: (tr, intr)]
      [
        //sense 1
        [ gram: [ gcode: (VP2A, VP15A) ]
          def: [ text: do clever tricks... ] ]
        [
          //sense 2
          [ gram: [ gcode: VP15B ]
            related: [ orth : conjure up ] ]
          [
            //sense 3
            [ form: [ pron: kən"dzU@(r) ]
              gram: [ gcode: VP17 ]
              usage: [ reg: formal ]
              def: [ text : appeal solemnly to... ] ]
            ...
          ]
        ]
      ]
    ]
  ]
}

```

Fig. 29. Overriding of values

//sense 1	[	form: [ orth: conjure hyph: con.jure pron: "kVndZ@r) ] gram: [ pos: v subc: (tr, intr) gcode: (VP2A, VP15A) ] def: [ text: do clever tricks... ]	]
//sense 2	[	form: [ orth: conjure hyph: con.jure pron: "kVndZ@r) ] gram: [ pos: v subc: (tr, intr) gcode: VP15B ] related: [ orth : conjure up ]	]
//sense 3	[	form: [ orth: conjure hyph: con.jure pron: k@n"dzU@r) ] gram: [ pos: v subc: (tr, intr) gcode: VP17 ] usage: [ reg: formal ] def: [ text : appeal solemnly to... ]	]

Fig. 30. Unfactored version of the feature structure in Fig. 29

## 4 Implementation: An object-oriented prototype

An implementation of the model described above presents difficulties because there exist no DBMSs based on features structures. The feature-based systems developed so far are designed for parsing natural language and are not intended to be used as general DBMSs. Therefore, they typically do not provide even standard database operations. They are furthermore usually restricted to handle only a few hundred grammar rules, and so even the largest systems are incapable of dealing with the large amounts of data that would be required for a dictionary.

We have already seen that existing general DBMSs, including relational and unnormalized systems, are too limited to handle lexical data.<sup>9</sup> However, a new generation of DBMSs which are *object-oriented* may provide the required expressiveness and flexibility. Object-oriented models allow for highly structured

---

<sup>9</sup>Some recently developed NF<sup>2</sup> relational models (Schek, et al., 1990) allow recursive nesting of relations, which could provide a more natural means to represent lexical data in a relational system. We have not explored this possibility.

objects by enabling the construction of new types, and in particular, recursive types, as well as providing complex built-in type constructors such as lists and sets. The underlying principle of the object-oriented approach is to eliminate computer-based concepts such as records and fields (the fundamental concepts in the relational model), and enable the user to deal with higher-level concepts that correspond more directly to the real world objects the database represents. Objects within the database, together with all of the attributes associated with them (and even operations and functions for manipulating these attributes) are considered as wholes, whereas in relational models, objects do not exist as wholes but are instead split across the various relations defined in the database.

A number of object-oriented DBMSs are currently operational (for instance, GemStone, GBASE, ONTOS --see Gardarin & Valduriez, 1990). Our implementation uses the O<sub>2</sub> system, which is outlined briefly in section 4.1. Section 4.2 demonstrates how the feature-based model for dictionaries is mapped into the O<sub>2</sub> data model. Section 4.3 describes the O<sub>2</sub> implementation.

#### 4.1 The O<sub>2</sub> system and model

O<sub>2</sub> is an object-oriented DBMS specifically designed for "new applications" such as CAD/CAM or editorial information systems and office automation (Deux *et al.*, 1991). The O<sub>2</sub> environment includes:

- an object-oriented "4th generation" programming language (O<sub>2</sub>C), which is an extension of C that enables database manipulation and user interface generation;
- a query language (O<sub>2</sub>Query), which is an extension of SQL that enables handling complex values and objects. The query language can be used independently and interactively, or it can be called from O<sub>2</sub>C;
- a user interface generator (O<sub>2</sub>Look), based on Motif and XWindows;
- a object-oriented programming environment (debugger, database browser, etc.).

An O<sub>2</sub> database (see Lecluse & Richard, 1989) consists of a set of *objects*, each of which consists of an identifier-value pair  $\langle i, v \rangle$ . The identifier for any object is unique. Values may be either null, atomic (integers, reals, strings, and booleans) or complex, in which case they are defined as follows:

- if  $n_1, \dots, n_p$  are attributes and  $x_1, \dots, x_p$  are values or identifiers, then  $tuple(n_1: x_1, \dots, n_p: x_p)$  is a value,
- if  $x_1, \dots, x_p$  are values or identifiers then  $set(x_1, \dots, x_p)$  and  $list(x_1, \dots, x_p)$  are values.

Note that we restrict the form of feature structures in our model to a *hierarchical normal form*. That is, in any feature structure  $F = [\phi_1, \dots, \phi_n]$ , only one  $\phi_i$ , let us say  $\phi_n = \{\psi_1, \dots, \psi_p\}$ , is a disjunction. This restriction is applied recursively to embedded feature structures. This scheme enables representing a feature structure as a tree in which factored information  $[\phi_1, \dots, \phi_{n-1}]$  at a given

level is associated with a node, and branches from that node correspond to the disjuncts  $\psi_1, \dots, \psi_p$ . Information associated with a node applies to the whole subtree rooted at that node. For example, the tree in Fig. 24 represents the feature structure for *abandon* given in Fig. 22. The representation of information as a tree mirrors the hierarchical structure and factoring of information in dictionaries.

For example, the following are objects:

```
<o0, tuple(name: "Fred", spouse: o1, children: set(o2, o3))>
<o1, tuple(name: "Mary", spouse: o0, children: set(o2, o3))>
<o2, tuple(name: "John", spouse: null, children: null)>
<o3, tuple(name: "Paul", spouse: null, children: null)>
```

Objects are grouped in *classes*, which correspond more or less to the traditional notion of abstract data types. Each class is defined by its name, the type of the value of its objects, and a set of *methods*, that is, the set of operations or functions that can be performed on objects of that class. Classes are organized in a class hierarchy, where subclasses automatically inherit methods defined for the superclass. Types for subclasses can be specialized according to a partial order relation (*subtyping*) among types.

## 4.2 Mapping feature structures into the O<sub>2</sub> model

The O<sub>2</sub> and feature structure models bear certain obvious similarities. Simple feature structures correspond to tuples in the O<sub>2</sub> system; features are analogous to attributes. For example, the feature structure

$$\begin{bmatrix} f: a \\ g: b \end{bmatrix}$$

can be translated into the O<sub>2</sub> object

```
<o1, tuple(f: "a", g: "b")>
```

Complex feature structures are translated into multiple objects in O<sub>2</sub>:

$$\begin{bmatrix} f: a \\ g: \begin{bmatrix} h: b \\ i: c \end{bmatrix} \end{bmatrix}$$

becomes the set of O<sub>2</sub> objects

```
<o1, tuple(f: "a", g: o2)>
<o2, tuple(h: "b", i: "c")>
```

Each feature structure type corresponds to an O<sub>2</sub> class. Since type schemas for feature structures specify all possible features that may appear, when implemented in O<sub>2</sub>, attributes corresponding to missing features will have null values.

Value disjunction in feature structures is implemented in O<sub>2</sub> using sets and lists. For example,

$$\left[ \begin{array}{l} f: a \\ g: \left\{ \begin{array}{l} h: b \\ i: c \\ \hline h: d \end{array} \right\} \end{array} \right]$$

is translated into

```
<o1, tuple(f: "a", g: set(o2, o3))>
<o2, tuple(h: "b", i: "c")>
<o3, tuple(h: "d", i: null)>
```

O<sub>2</sub> does not provide a built-in mechanism to handle general disjunction. However, general disjunction can be implemented through recursive types. Close examination of the feature structure model above shows that every node in the hierarchical normal form tree (see Fig. 24) has the same type. In O<sub>2</sub>, this is implemented by introducing an additional recursive attribute DISJ in tuples. For example, if type T for some feature structure is

$$\left[ \begin{array}{l} f: string \\ g: string \\ h: string \end{array} \right]$$

the class definition for T in O<sub>2</sub> is<sup>10</sup>

```
class T
  type tuple (f: string,
              g: string,
              h: string,
              DISJ: set (T))
end
```

Then, the feature structure of type T:

$$\left[ \begin{array}{l} f: a \\ \left\{ \begin{array}{l} \left[ \begin{array}{l} g: b \\ h: c \\ \hline g: d \\ h: e \end{array} \right] \end{array} \right\} \end{array} \right]$$

is implemented as

---

<sup>10</sup>Note that because of the way we implement general disjunction in O<sub>2</sub>, it is necessary to indicate the type of disjunction (set or list) for each feature-structure type at the time it is declared. This is not required in the feature-based model.

```
<o1, tuple(f: "a", g: null, h: null, DISJ: set(o2, o3))>
<o2, tuple(f: null, g: "b", h: "c", DISJ: null)>
<o3, tuple(f: null, g: "d", h: "e", DISJ: null)>
```

Note that the unfactor operator has to be programmed as a method, since general disjunction is not a built-in mechanism in O<sub>2</sub>. The unification operation on which unfactor relies (see section 3.5) is also not built-in.

Fig. 31 shows a few simplified class definitions appropriate for a lexical database. Fig. 32 shows in graphic form the O<sub>2</sub> implementation for the entry *abandon* from the LDOCE. Each box in the figure corresponds to an O<sub>2</sub> object.

```
class Entry
  type tuple (form: list (Form),
              gram: list (Gram),
              def: list (Def),
              ...
              DISJ: list (Entry))
  end;

class Form
  type tuple (orth: list (string),
              pron: list (string),
              hyph: list (string),
              geo: list (string),
              DISJ: list (Form))
  end;

class Gram
  type tuple (pos: list (string),
              subc: list (string),
              gend: list (string),
              numb: list (string),
              DISJ: list (Gram))
  end;
  ...
```

Fig. 31. O<sub>2</sub> class definitions for a lexical database

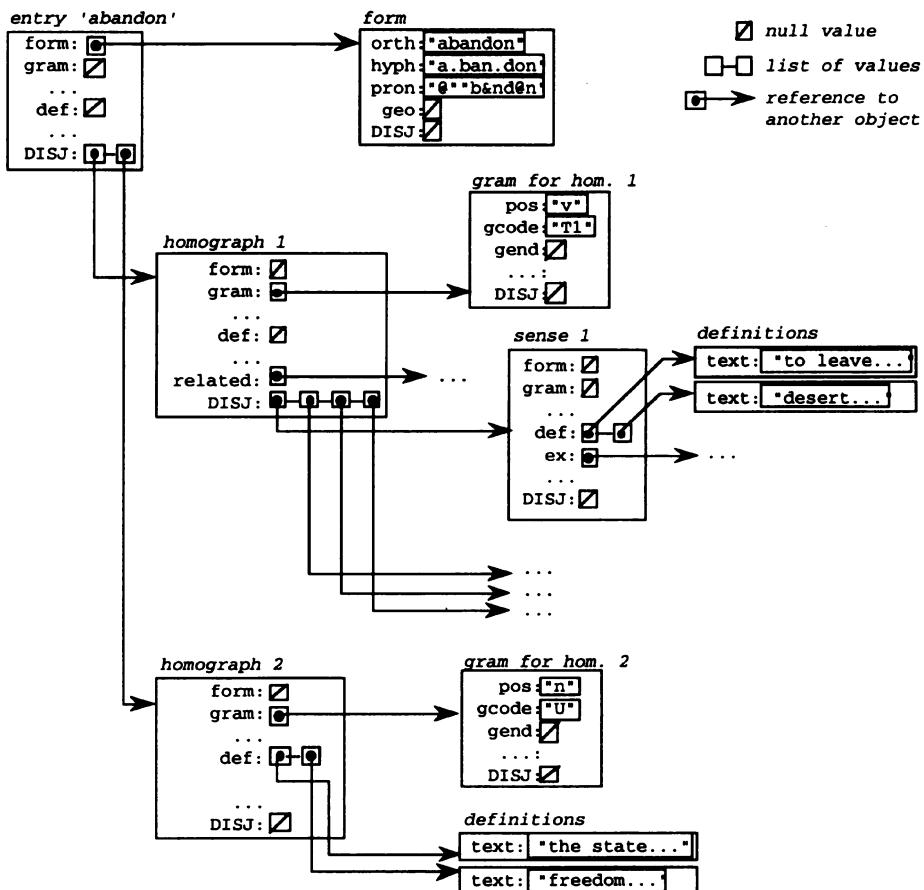


Fig. 32. Representation of the entry 'abandon' in O2

## 4.3 Implementation

We have implemented a lexical database in a prototype version of O2. The different classes were defined according to the schema described in the preceding section, and the methods were programmed in O2C.

### 4.3.1 Creation of the lexical database

The dictionary we have used is the *Zyzomys*, published by Hachette and distributed on CD-ROM. The *Zyzomys* is encoded with mixed markup, including both procedural markup (for example, /IT indicates italicized text, /RO indicates roman, etc.) and descriptive markup (for example, /DP and /FP delimit the

phonetic transcription, /ME marks the orthographic form) (see sections 2.1.1 and 2.1.2 and Fig. 33).

**gin** [dZIn] n. m. Eau-de-vie de grain aromatisée au genièvre, fabriquée notam. en G-B. - Mot angl. *gin*, adapt. du neerl. *jenever*, =genièvre=.

/MDGIN/FD /MEgin /FE /DPdZin/FP /GE/BGn. m./GG438/GB/FG Eau-de-vie de grain aromatisée au genièvre, fabriquée  
/BGnotam./GG445/GB en /BGG.-B./GG262/GB - Mot  
/BGang1./GG035/GB /ITgin/RO, adapt. du /BGneerl. /GG433/GB  
/ITjenever/RO, =genièvre=.

Fig. 33. The entry for 'gin' from the *Zyzomys* and its original encoding

The first step was to analyze the entire dictionary in order to isolate the different logical fields composing each entry, and organize them according to the model defined above. The results of this analysis were encoded in SGML according to a preliminary version of the guidelines for encoding monolingual dictionaries, which we developed while working within the Text Encoding Initiative (Ide, Véronis, Warwick-Armstrong & Calzolari, in press).

Each feature of our model (for example, *form*, *gram*, etc.) corresponds to an SGML element (<form>...</form>, <gram>...</gram>, etc.--see Fig. 34), as well as to an O<sub>2</sub> class (Form, Gram, etc.).<sup>11</sup> The dictionary is translated from its SGML format by means of a recursive descent procedure: each time a new opening tag (for example, <form>) is encountered, a method in the corresponding O<sub>2</sub> class (for example, Form) is triggered, which translates the content of the element concerned into an O<sub>2</sub> value or object.

A few problems were encountered in the process of creating the database, mainly because the version of O<sub>2</sub> used in the experiments was a prototype. In our model, each entry in the dictionary is represented by a tree structure consisting of potentially several dozens of objects. The object manager of our prototype version of O<sub>2</sub> could not handle the hundreds of thousands of objects corresponding to the approximately 50,000 entries of the *Zyzomys*. We therefore limited our experiments to 600 entries. This problem is completely resolved in the commercial version of O<sub>2</sub>, which is now available.

<sup>11</sup>Etymology is not included in our database.

form:	[orth: gin]
	[pron: dzIn]
gram:	[pos: n]
	[gen: m]
def:	[text: Eau-de-vie de grain aromatisée au genièvre, fabriquée notam. en G.-B.]

(a) feature structure

```

<entry>
  <form>
    <orth>gin</orth>
    <pron>dzIn</pron>
  </form>
  <gram>
    <pos>n</pos>
    <gen>m</gen>
  </gram>
  <def>
    <text>Eau-de-vie de grain aromatisée au
          genièvre, fabriquée notam. en G.-B.</text>
  </def>
</entry>

```

(b) SGML encoding

Fig. 34. Representations of 'gin' from the *Zyzomys*.

Null values caused a second problem. In both the prototype and commercial versions of O<sub>2</sub>, null values are explicitly stored, which, especially for the sparse data in lexical entries, is very space-consuming. However, it is conceivable that future versions of O<sub>2</sub> or other object-oriented DBMSs could solve this problem by not representing null values internally.

#### 4.3.2 Querying the database

A typical query asks to extract all entries (or parts of entries) which have certain attribute values (for example, the query for the *LDOCE* given as an example in section 2.2.1, intended to extract all examples for countable nouns whose definitions start with "the state..."--similar queries could be applied to the *Zyzomys*). When the user queries the database, he or she does not know *a priori* how the target definitions are factored (or even if they are all factored in the same way) and therefore at what level certain attributes such as gram.pos and def.text appear.

We have programmed a query interface in O<sub>2</sub>C, which enables the user to formulate queries in an unfactored, "flat" format and determines that attribute values appear at appropriate places within the tree. The retrieval process involves two steps. First, an index for each atomic attribute (gram.pos, gram.gcode, def.text, etc.) enables retrieving all entries with a given value for that attribute. In the *LDOCE* example, the indexes will enable retrieving all entries containing all of the attribute-value pairs in the query (gram.pos : "n", gram.gcode : "U", def.text : "state"). However, because there is no indication in the indexes of

where the attribute-value pairs appear in the entry trees, this first step will also retrieve, for example, entries in which there exist both a noun and a verb homograph, and where "state" appears in the definition text for the verb homograph. Similarly, it will retrieve entries where "n" has been overridden at a lower level and the definition text containing "state" appears at this lower level. Therefore, a second step is necessary, to recursively traverse the trees of the entries retrieved in the first step, and retain only those which match the query exactly.

This solution is not completely satisfactory, since it by-passes the query language and therefore does not take advantage of its features. Ideally, the database query language should include an operator enabling the traversal of recursive structures. Because O<sub>2</sub>Query does not allow the definition of new operators, we are working on an extension to the query language LIFOO (developed in part by one of the authors as a functional query language for O<sub>2</sub>-- see Le Maitre & Boucelma, *in press*) which is specifically constructed to support the definition of new operators (Boucelma & Le Maitre, 1991).

## 5 Conclusion

In this paper we show that previously applied data models are inadequate for lexical databases. In particular, we show that relational data models, including normalized models which allow the nesting of attributes, cannot capture the structural properties of lexical information. We propose an alternative feature-based model for lexical databases, which departs from previously proposed models in significant ways. In particular, it allows for a full representation of sense nesting and defines an inheritance mechanism that enables the elimination of redundant information. The model provides flexibility which seems able to handle the varying structures of different monolingual dictionaries. Our model may therefore be applicable to a diversity of uses of electronic dictionaries, ranging from research to publication.

We also show how the feature-based model can be implemented in an object-oriented DBMS, and demonstrate that feature structures map readily to an object-oriented data model. Our work suggests that the development of a feature-based DBMS, including built-in mechanisms for disjunction, unification, generalization, etc., is desirable. Such feature-based DBMSs could have applications far beyond the representation of lexical data.

A number of open problems remain for fully specifying the structure and elements of lexical databases. For example, we have not addressed the problems of phrasal elements (such as discontinuous verb phrases and cross-reference phrases embedded in definition or example text), etymologies (which are themselves complex structured text), etc. Further, it is necessary to test our model across a wide range of monolingual dictionaries in order to ascertain, first, its generality and, second, the exact scope and nature of remaining difficulties.

## Acknowledgments

The present research has been partially funded by the GRECO-PRC Communication Homme-Machine of the French Ministry of Research and Technology, U.S.-French NSF/CNRS grant INT-9016554 for collaborative research, and U.S. NSF RUI grant IRI-9108363. The authors would like to acknowledge the GIP-Altaïr for making available a prototype version of the O<sub>2</sub> DBMS (contract GIP-Altaïr #88-5), and Collins Publishers, Hachette, Longman Group, and Oxford University Press for making their data available for research within the project. The authors would also like to thank Christine Tribocky, Régis Voillaume, and Christiane Fantozzo for their work on the implementation, Jean-Michel Ombrouck for his pre-processing of the Hachette *Zyzomys*, Mary Neff for providing the IBM LDB example, and Frank Tompa for his valuable comments on an earlier draft of this paper.

## References

- Abiteboul, S., & Bidoit, N. (1984). Non first normal form relations to represent hierarchically organized data. *Proceedings of the ACM SIGACT/SIGMOD Symposium on Principles of Database Systems*. Waterloo, Ontario, 191-200.
- Amsler, R. A. (1980). *The structure of the Merriam-Webster Pocket Dictionary*. Doctoral dissertation, University of Texas at Austin.
- Amsler, R. A., & Tompa, F. W. (1988). An SGML-based standard for English monolingual dictionaries. *Proceedings of the 4th Annual Conference of the UW Centre for the New Oxford English Dictionary*. Waterloo, Ontario, 61-80.
- Bancilhon, F., Fortin, D., Gamersan, S., D., & Verroust, A. (1983). VERSO: A relational backend database machine. In D. K. Hsiao (Ed.), *Advanced Database Machine Architecture*. Englewood Cliffs: Prentice Hall.
- Boguraev, B., & Neff, M. S. (in press). From machine readable dictionaries to lexical databases. *International Journal of Lexicography*.
- Boucelma, O., & Le Maitre, J. (1991). An extensible functional query language for an object-oriented database system. In C. Delobel, M. Kifer, Y. Masunaga (Eds.), *Deductive and Object-Oriented Databases*. Lecture Notes in Computer Science, Berlin: Springer Verlag.
- Brustkern, J., & Hess K. (1982). The BONNLEX lexicon system. In J. Goetschalckx & L. Rolling (Eds.), *Lexicography in the Electronic Age*. Amsterdam: North-Holland.
- Byrd, R. J., Calzolari, N., Chodorow, M. S., Klavans, J. L., Neff, M. S., & Rizk, O. Tools and methods for computational linguistics. *Computational Linguistics*, 13(3/4), 219-240.
- Calzolari, N. (1984). Detecting patterns in a lexical data base. *Proceedings of the 10th International Conference on Computational Linguistics, COLING'84*. Stanford, California, 170-173.

- Calzolari, N., Peters, C., & Roventini, A. (1990). *Computational Model of the Dictionary Entry*. (ACQUILEX Preliminary Report, Esprit Basic Research Action No. 3030). Pisa, Italy: Istituto di Linguistica Computazionale.
- Coombs, J. H., Renear, A. H., & DeRose, S. J. (1987). Markup systems and the future of scholarly text processing. *Communications of the ACM*, 30(11), 933-47.
- Deux, O. et al. (1991). The O<sub>2</sub> System. *Communications of the ACM*, 34(10), 34-48.
- Gardarin, G., & Valduriez, P. (1990). *SGBD Avancés, Bases de Données Objets, Déductives, Réparties*. Paris: Eyrolles.
- Gonnet, G., & Tompa, F. W. (1987). Mind your grammar: a new approach to modelling text. *Proceedings of the 13th Conference on Very Large Data Bases, VLDB'87*. Brighton, England, 339-346.
- Gonnet, G., Baeza-Yates, R. A., & Snider, T. (1991). *Lexicographical indices for text: Inverted files vs. PAT trees* (Technical report OED-91-01). Waterloo, Ontario: UW Centre for the New Oxford English Dictionary and Text Research.
- Ide, N., Véronis, J., Warwick-Armstrong, S., & Calzolari, N. (in press). Principles for Encoding machine readable dictionaries. *Proceedings of the Fifth EURALEX International Congress, EURALEX'92*. Tempere, Finland.
- Kaplan, R. and & Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*. Cambridge, Massachussets: MIT Press.
- Karttunen, L. (1984). Features and values. *Proceedings of the 10th International Conference on Computational Linguistics, COLING'84*. Stanford, California, 28-33.
- Kay, M. (1985). Parsing in functional unification grammar. In D.R. Dowty, L. Karttunen, & A. M. Zwicky (eds.). *Naturality Press*.
- Kipfer, B. A. (1983). Computer applications in lexicography: a bibliography. *Dictionaries: Journal of Dictionary Society of North America*, 4, 202-237.
- Klavans, J., Chodorow, M., & Wacholder, N. (1990). From dictionary to knowledge base via taxonomy. *Proceedings of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary*. Waterloo, Ontario, 110-132.
- Landau, S. I. (1984). *Dictionaries: The Art and Craft of Lexicography*. New York: Scribner Press.
- Le Maitre, J., & Boucelma, O. (in press). LIFOO, un langage fonctionnel de requêtes pour bases de données avancées. *Technique et Science Informatiques*.
- Lecluse, C., & Richard, P. (1989). The O<sub>2</sub> database programming language. *Proceedings of the 15th Conference on Very Large Data Bases, VLDB'87*. Amsterdam, 411-422.
- Markowitz, J., Ahlswede, T., & Evens, M. (1986). Semantically significant patterns in dictionary definitions. *Proceedings of the 24th Annual Conference of the Association for Computational Linguistics*. New York, 112-119.
- Nakamura, J., & Nagao, M. (1988). Extraction of semantic information from an ordinary English dictionary and its evaluation. *Proceedings of the 12th*

- International Conference on Computational Linguistics, COLING'88.*  
Budapest, Hungary, 459-464.
- Neff, M. S., Byrd, R. J., & Rizk, O. A. (1988). Creating and querying lexical databases. *Proceedings of the Association for Computational Linguistics Second Applied Conference on Natural Language Processing*. Austin, Texas, 84-92.
- Pistor, P., & Traunmueller, R. (1986). A database language for sets, lists and tables. *Information Systems*, 11(4), 323-336.
- Pollard, C., & Sag, I. A. (1987). *Information-based Syntax and Semantics*. CSLI Lecture Notes Series, Chicago: University of Chicago Press.
- Roth, M. A., Korth, H. F., & Silberschatz, A. (1988). Extended algebra and calculus for nested relational databases. *ACM Transactions on Database Systems*, 13(4), 389-417.
- Schek, H.-J., Paul, H.-B., Scholl, M.H., & Weikum, G. (1990). The DASDBS project: objectives, experiences, and future prospects. *IEEE Transactions on Knowledge and Data Engineering*, 2(1), 25-42.
- Shieber, S.M. (1986). *An Introduction to Unification-based Approaches to Grammar*. CSLI Lecture Notes Series, Chicago: University of Chicago Press.
- Sinclair, J. M. (1987). *An Account of the COBUILD Project*. London: Collins ELT.
- Sperberg-McQueen, M., & Burnard, L. (1990). *Guidelines for the encoding and interchange of machine-readable texts, Draft, Version 0.0*. ACH, ACL, and ALLC.
- The DANLEX Group (1987). Descriptive tools for electronic processing of dictionary data. *Lexicographica, Series Maior*. Tübingen: Niemeyer.
- Tompa, F. W. (1989). What is tagged text? *Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary*. Oxford, England, 81-93.
- Véronis, J., & Ide, N., M. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90*. Helsinki, Finland, 2, 389-394.
- Wilks, Y., Fass D., Guo, C., MacDonald, J., Plate, T., & Slator, B. (1990). Providing Machine Tractable Dictionary Tools. *Machine Translation*, 5, 99-154.

# Construction-Based MT Lexicons

Lori Levin  
Sergei Nirenburg  
Carnegie Mellon University  
*e-mail:* {*lsl, sergei*}@nl.cs.cmu.edu

## Abstract

This paper presents a novel view of the boundary between the generalizable and the idiosyncratic in MT lexicons. We argue that the domain of the idiosyncratic should, in fact, be broader than in most current approaches. While at present most MT systems involve phrasal lexicons, these typically contain terminology from a particular field. In order to facilitate naturalness of translation, specifically, to carry the level of “conventionality” of meaning expression across languages, it becomes necessary to use the concept of a grammatical construction, a (possibly, discontiguous) syntactic structure or productive syntactic pattern whose meaning it is often impossible to derive solely based on the meanings of its components. Identification of constructions allows an MT system to select the most appropriate conventional way of expressing a meaning from among the available ways. After discussing the notion of construction, we suggest the format for a construction lexicon for a knowledge-based MT system.

## 1 Introduction

Source text analyzers and target text generators in all rule-based MT systems rely on a variety of knowledge sources, centrally including grammars and lexicons. Grammar rules vary in generality (productivity) in that they can apply to very broad class of phenomena, such as, for instance, all adjectives, or to a very narrow class, even a single lexical unit. Information in lexicons typically relates to a single lexical entry (which, however, can be phrasal), but modern computational lexicons are often organized in such a way that some information applies to (often, broad) classes of entries.

There are many reasons to attempt to write grammar rules in the most general manner—the more generally applicable the rules, the fewer rules need to be written; the smaller the set of rules (of a given complexity) can be found to be sufficient for a particular task, the more elegant the solution, etc. In the area of the lexicon, the ideal of generalizability and productivity is to devise simple entries which, when used as data by a set of syntactic and semantic analysis operations, regularly yield predictable results in a compositional manner. To be maximally general, much of the information in lexical entries should be inherited based on class membership or should be predictable from general principles.

However, experience with NLP applications shows that the pursuit of generalization promises only limited success. In a multitude of routine cases it becomes difficult

to use general rules. This leads to the necessity of directly recording (usually, in the lexicon) information about how to process small classes of phenomena which could not be covered by general rules. An important goal for developers of NLP systems is, thus, to find the correct balance between what can be processed on general principles and what is idiosyncratic in language, what we can calculate and what we have to know literally, what is compositional and what is conventional. In other words, the decision as to what to put into a set of general rules and what to store in a static knowledge base such as the lexicon becomes a crucial early decision in designing computational-linguistic theories and applications.

The trade-off between generality and idiosyncraticity is complicated by the possibility of the following compromise. If direct generalizations cannot be made, there may still be a possibility that the apparent variability in grammar rules and lexicon data can be accounted for by *parameterization*: there may exist a set of universal parameters that would explain the differences among various phenomena in terms of the difference in particular parameter settings. This is better than dealing with ungeneralized material. But the search for a set of universal parameters, however important, does not, in our opinion, hold a very bright promise from the standpoint of coverage.

In this paper we discuss a set of phenomena, *constructions*, which do not lend themselves to either a generalized or parameterized treatment, while at the same time differing from the predominantly, terminological material typically appearing in the phrasal parts of NLP lexicons.

This paper does not address the issue of how constructions can be actually used in the process of extracting meaning from texts, though in our long-term research program this is a central issue. In a nutshell, our approach to natural language analysis is built on a constellation of “microtheories” of particular language phenomena, united at the level of system architecture and representation language. The microtheory rules for calculating various components of meaning use a variety of diverse clues—morphological, syntactic, lexical—in their input conditions. We believe that knowledge about constructions—significant combinations of syntactic structures, lexical items and other morphemes—contributes greatly to the expressive power of such rules.

## 2 Idiosyncractic Phenomena in MT

The concept of construction was revived by Charles Fillmore, Paul Kay and their students (e.g., Fillmore et al., 1988, Fillmore and Kay, unpublished). According to this approach, the specification of a construction can include syntactic, semantic, and pragmatic information, but the semantics and/or pragmatics can be different from the compositional semantics and/or pragmatics normally associated with the same structure by productive rules. Furthermore, many constructions, such as those below, violate do not conform to general syntactic rules. Constructions are, therefore, like words in that they have to be learned separately as integral facts about language. At the same time, constructions are not the same as frozen idioms; they can be productive grammatical patterns, many of whose properties are predictable from general principles. The following are some examples of English and Russian constructions. The English examples are taken from Fillmore et al. (1988).

- (1) a. What with the kids off to school and all.  
     b. Why not fix it yourself?  
     c. Him be a doctor?  
     d. What do you say we stop here?  
     e. It's time you brushed your teeth.  
     f. One more and I'll leave.  
     g. No writing on the walls!
- (2) a. Chto ni       govori  
         what CLITIC say-IMPERATIVE  
         a           matematika    interesnyj   predmet  
         CONJ   mathematics   interesting   object  
         “Whatever you might say, mathematics is an interesting subject.”
- b. Chto zhe       eto               ty  
         What CLITIC this-NEUTER you-NOMINATIVE  
         Ivan Ivanovich, zabyl o  
         Ivan Ivanovich, forgot about  
         nashem dogovore?  
         our       agreement-PREPOSITIONAL  
         “How come you forgot about our agreement, Ivan Ivanovich?”
- c. Kuda           nam           do nix.  
         Where-DIRECTION we-DATIVE to they-GENITIVE  
         “How can we compete/compare with them.”

Constructions with non-compositional semantics and pragmatics are not rare exceptions to rules. They co-exist with the basic lexis and grammar of language and in many cases present the most typical, unmarked way for expression of a particular meaning. Thus, the rules governing the use of constructions and the “regular” rules must be made to co-exist in any application, as they are equally important for associating semantic and pragmatic effects with utterances.

## 2.1 Conventionality and Constructional Divergences in MT

The *conventionality* of constructions is one of their defining characteristics. Constructions can be understood as conventional in two ways. First, meanings are associated with constructions by convention. That is, many constructions, similarly to separate words, have an arbitrary (non-compositional, non-iconic) association with their meanings. Among the types of meaning often associated with constructions are aspect,

time/tense, modality, evidentiality, speaker attitude, speech act, conditionality, comparison, causality, rhetorical relations, etc. The microtheory approach to these phenomena makes abundant use of constructions as sources of information about what meanings are present in a sentence.

The other sense of conventionality concerns the typical default ways of expressing meanings in language, which may have been grammaticalized as an arbitrary form-meaning relationship. It should not be necessary to involve inference processes for the analyzer to arrive at the intended meaning. For example, (3)a is a conventional way of requesting that someone pass the salt, whereas (3)b is not a conventional request. We propose different treatments for (3)a and (3)b. The former matches an entry in a *construction lexicon* and does not require inference. The latter requires inference in order to be interpreted as a request. ((3)a could also involve inference based on felicity conditions for requests, but its conventionality eliminates the need for inference.) All constructions in our lexicon are conventional in the first sense (that is, non-compositional) but not all of them are conventional in the second sense. That is, in some cases a different realization of the same meaning could be less marked.

- (3) a. Can you pass the salt?  
b. Gee, this food is bland.

Conventionality is an important factor in translation. Thus, conventional expressions of a meaning in the source language should be translated into similarly conventional expressions of the same meaning in the target language. For example, the Japanese sentences in (4) are conventional expressions of the modal meaning of obligation and should therefore be translated into a conventional expression of obligation in English such as *You should go* or *You'd better go*. Literal translations of these sentences into English, (*Not going won't do* and *The alternative that (you) went is good*), while understandable, are not appropriate translations, due to their low rating on the conventionality scale.

- (4) a. Itta hoo ga ii.  
go-PAST alternative NOM good  
Literally: The alternative that (you) went is good.
- b. Ikanakute wa ikenai.  
go-NEG-GERUND TOP won't do  
Literally: Not going won't do.
- c. You'd better go./ You should go.

In translation, source text elements can vary in their degree of conventionality in expressing the meanings they carry. It is natural to require that the translation has the conventionality level closest to that of the source text. The requirement of maintaining conventionality levels in translation is a source of *constructional divergences*, instances (such as (4)) in which, in order to retain the level of conventionality, a target language passage is selected with a syntactic structure very different from that of the corresponding source passage. Further examples of constructional divergences are shown in (5)-(7).

- (5) a. Ich esse gern.  
          I eat likelyly  
          Literally: I eat likelyly
- b. I like eating. (Dorr, 1992)
- (6) a. Juan suele                 ir             a casa.  
          Juan be-in-the-habit-of go-INF to house  
          Literally: Juan tends to go home.
- b. John usually goes home. (Dorr, 1992)
- (7) a. On zagovoril.  
          he speak-INCHOATIVE  
          ‘He started to speak’
- b. He started to read a book.

Examples like these, when discussed in the literature, have led to the impression that constructional divergences arise from a fairly limited set of correspondences (such as a syntactic head in one language corresponding to a modifier in another language). The Japanese examples in (4) and the Russian examples in (2) above show that this is not the case; the structures involved in constructional divergences can be radically different and, for the most part, do not appear to follow from predictable or regular correspondences between source and target languages.

## 2.2 What is the Impact of Constructions on MT?

Exploiting constructions seems to be the only way toward guaranteeing the production of conventional (“colloquial”) rather than literal translations, although history of MT research shows neglect of the issue. Recent interest in MT divergences centers on linking and lexicalization divergences, which are less of a problem to solve, while largely ignoring the problem of constructions. It is clear that the expressive power of MT systems will grow significantly with the introduction of large construction lexicons.

## 3 Treatment of Constructional Divergences

A literal translation (that is, a translation which seeks to preserve in the target text the exact word and structure choice in the source text), even when formally possible, seldom succeeds in preserving the level of conventionality of the input text. Often a marked, unusual, compositional realization of the same meaning is obtained as illustrated in example (4). We believe that the treatment of constructional divergences in the lexicon is possible both for transfer-oriented and interlingual MT systems. Since our prior work in MT has centered mostly on the latter approach, our further discussion will be devoted to the ways of introducing the treatment of constructions into a lexicon structure similar to that used in some of our MT projects (e.g., DIANA, KBMT-89 or Mikrokosmos).

The text meaning representation produced for a construction will not, in the general case, be isomorphic to the syntactic representations of the source and target texts (see Nirenburg and Levin, 1992 for a discussion). This conclusion is further corroborated by the following consideration. Our theory of text meaning distinguishes between core semantic dependency statements (which we will call “the propositional content”) and additional semantic information that covers meanings such as aspect, time/tense, modality, evidentiality, speaker attitude, speech act, conditionality, comparison, rhetorical relations and others (which we will, for the sake of symmetry, call “non-propositional content” and which we represent as feature-value sets scoping over predicate-argument structures). The means to express these phenomena are among the most divergent among languages and at the same time are not readily parameterizable or generalizable.

Indeed, constructional divergences cannot be accounted for with a few parameters like head switching or locus of linking inside a semantic structure (Dorr, 1992). In our terms, constructions are used as a means of conventional, language-specific encoding of language-independent meaning. For example, the fact that the Japanese “Sentence-past *hoo ga ii*” conventionally encodes the meaning of suggestion or obligation is as much a part of the lexicon of Japanese as any definition of a word meaning. We introduce a *construction lexicon* as a repository of knowledge supporting both the treatment idiosyncratic, non-compositional constructions and the compositional realization of a variety of propositional meanings.

Before suggesting a possible structure of a construction lexicon entry, we would like to clarify a potential misunderstanding with respect to the definition of constructional divergences. It is important to distinguish constructional divergences from other circumstances that call for a target language translation to be structurally different from the source. For example, lexical gaps are typically treated in translation through optional, usually inferentially-produced paraphrases. For example, because there is a lexical gap for *afford* in Russian, a sentence like (8)a must be rendered in Russian as the translation of a sentence such as (8)b or (8)c. Examples in (4)-(7) are different in kind from the ones in (8) because in a computational implementation they should not involve paraphrasing through inference making but rather a look-up in a lexicon of conventional constructions (see below). Note that there are some indications that lexical gaps and constructional divergences form a scale rather than a dichotomy.

- (8) a. John can't afford a BMW.  
b. John does not have enough money to buy a BMW.  
c. John cannot allow himself to buy a BMW.

## 4 Some Examples

This section contains two examples of treatment of constructions in the framework of an interlingual MT system. In particular, it illustrates the lexicon entry structure and the interlingua (the text meaning representation, or TMR). The examples illustrate the use of constructions as units of analysis alongside words.

The examples also illustrate our treatment (or, rather, in our model, lack of the need for the treatment) of MT divergences—situations in which a source language sentence

and its target language translation differ significantly in syntactic structure, syntactic category, or predicate-argument structure. No special mechanisms are needed to treat MT divergences in our model, as source and target language sentences are not expected to be isomorphic to the TMR or to each other. All that is needed in order to translate a sentence involving a divergence are source and target language lexical entries of the sort illustrated here that map different syntactic structures onto the same TMR.

For each example, we list a TMR, a source language syntactic structure and a target language syntactic structure. The languages used for illustration are English, Russian, and Japanese. (Since the system is symmetrical, we do not identify which is the source language and which is the target language in each example.) It should be obvious that the source and target language sentences can produce the same TMR even if their syntactic structures are not isomorphic.

The TMR structure consists of clauses which roughly correspond to the “who did what to whom” component of meaning but also includes information about speech acts, speaker attitudes, indices of the speech situation, and stylistic factors as well as relations (e.g., temporal ones) among any of the above, and other elements.

The examples also include the relevant zones of the source and target language lexical entries (namely, syntax and mapping to TMR). (This lexicon format is discussed in some detail by Meyer et al. 1990.) The first zone (*syn-struc*) specifies an LFG-style syntactic subcategorization frame (Bresnan, 1982) of a predicate including which grammatical functions (subject, object, complement, etc.) the predicate must appear with and any requirements the predicate has of those functions (case, syntactic category, specific lexical items, etc.).

The second lexical entry zone that we illustrate (*sem*) specifies the portion of TMR that is associated with the lexical item in question and how the components of the TMR correspond to the components of the syntactic zone. We have chosen examples in which the TMR is not isomorphic to the syntactic zone. In most of the examples, a complements of the lexical item heads the associated TMR. In these cases, the syntactic head of the sentence corresponds to a scope-taking operator or a simple feature-value pair in TMR.

## 4.1 Treating Speech Act Constructions

Consider the sentences in (9), which constitute conventional ways of making requests in Japanese and English. The TMR for both the English and the Japanese phrase is represented in Figure 1.

The TMR indicates that the speaker is performing the speech act of requesting (*speech-act-1*), that the request is that the hearer buy a book (*clause-1*), that the buying will occur after the time of speech (*relation-1*), and that some time after the buying (*relation-3*), the book will belong to the speaker (*relation-2*). The syntactic feature structures of the sentences are illustrated in Figure 2. The constituent structure of the Japanese sentence is presumably mono-clausal, but corresponds to the bi-clausal feature structure shown here. It is also possible that the feature structure should be tri-clausal, depending on the analysis of the potential morpheme. See Matsumoto (1992) for a recent discussion of these issues.

---

```

clause-1
  head:      buy-1
  agent:     *hearer*
  theme:     book-1

aspect:
  phase:    none
  duration: momentary
  iteration: single

speech-act-1
  type:      request-action
  scope:     clause-1
  speaker:   *speaker*
  hearer:    *hearer*

relation- 1
  type:      temporal-before
  from:     time-of-speech
  to:       time-of(clause-1)

relation-2
  type:      possession
  from:     *speaker*
  to:       book-1

relation-3
  type:      temporal-before
  from:     time-of(clause-1)
  to:       time-of(relation-2)

```

Figure 1. TMR for the Sentences in Example (9).

---

- (9) a. Hon o katte  
     book OBJ buy-GERUND  
     moraemasen                                 ka?  
     receive-POTENTIAL-FORMAL-NEG QUEST  
     “Can I receive the favor of you buying a book for me?”
- b. Can you buy me a book?

Figures 3, 4 and 5 show some of the the lexicon entries involved in building the syntactic structures shown in Figure 2 and the TMR shown in Figure 1.

The entries specify a) portions of the syntactic structure in which the lexical unit in question will appear, b) the meaning of the lexical unit, which could be viewed as “canned” portions of the TMR and c) correspondences between these two kinds of structure. The entries for *can* and *moraau* (receive) are examples of construction lexicon entries. They contain information that is specific to the use of these words in making requests such as the subject being *you* in English and the tense being non-past in

---

### English Syntactic Structure

PREDICATE	can
SUBJECT	you
COMPLEMENT	
PREDICATE	buy
SUBJECT	you
OBJECT	me
OBJECT2	book

### Japanese Syntactic Structure

PREDICATE	morau (receive)
MOOD	potential, interrogative
TENSE	non-past
OBJECT	hon (book)
SUBJECT	pronoun (speaker)
COMPLEMENT	
PREDICATE	kau (buy)
SUBJECT	pronoun (hearer)
OBJECT	hon (book)

Figure 2. Syntactic Structures for the Sentences in Example (10).

CAN:

```
Syn-Struc: predicate: [0] can
           subject: [1]
                     root: you
           complement: [2]
                     subject: [1]

Sem:      clause [3]
           head: meaning-of ([2])
           speech-act [4]
             type: request-action
             scope: [3]
             speaker: *speaker*
             hearer: *hearer*
           relation [5]
             type: temporal-before
             from: time-of-speech
             to:   time-of([3])
```

Figure 3. Construction Lexicon Entry for a Request in English.

---

Japanese. *Can* and *morau* have other lexical entries as well for their other senses and constructional uses. There are also many other construction lexicon entries for other ways of making requests in both languages.

---

BUY

```
Syn-Struc: predicate: [0] buy
           subject: [1]
           object: [2]
           object-2: [3]

Sem:      clause [4]
           head: meaning-of([0])
           agent: meaning-of ([1])
           theme: meaning-of ([3])
           relation [5]
           type: possession
           from: meaning-of([2])
           to: meaning-of([3])
           relation [6]
           type: temporal-before
           from: time-of([4])
           to: time-of([5])
```

Figure 4. Lexicon entry for English “buy.”

MORAU

```
Syn-Struc: predicate: [0] morau
           tense: non-past
           mood: potential
           subject: [1]
           root: speaker
           object: [2]
           object-2: [3]
           root: hearer
           complement: [4]
           inflection: gerund
           subject: [3]
           object: [2]

Sem:      clause [5]
           head: meaning-of ([4])
           speech-act [6]
           type: request-action
           scope: [5]
           speaker: *speaker*
           hearer: *hearer*
           relation [7]
           type: temporal-before
           from: time-of-speech
           to: time-of ([5])
```

Figure 5. Construction Lexicon Entry for a Request in Japanese.

---

Correspondences between syntactic elements in the *syn-struc* zone and elements of TMR in the *sem* zone are indicated by co-indexing. *Meaning-of* (*x*) is a function whose value is the TMR corresponding to the feature structure with index *x*. Notice that in the lexical entries for *can* and *morau* the complement of the *syn-struc* zone is co-indexed with the head of the main clause in the *sem* zone. In building a TMR for the English sentence in (9), this means that the *meaning-of* the syntactic clause headed by *buy*, given in the *sem* zone of Figure 4, will become the head of the main clause of the TMR in Figure 1. In other words, *buy* (or *kau*) conveys the main semantic content of the sentence. *Can* and *morau* serve only to trigger the speech-act *request-action* in the TMR.

## 4.2 Treating Modality

Our second example involves constructions that illustrate the modality of obligation in Japanese, Russian, and English, as shown in (10). The TMR corresponding to these sentences is shown in Figure 6.

```

clause-1
  head:          go-1
  agent:         *hearer*
  destination:   *unknown*
  aspect:
    phase:        none
    duration:     *unspecified*
    iteration:    single

attitude- 1
  type:          deontic
  value:         0.8-1.0
  scope:         clause-1
  attributed-to: *speaker*
  time:          time-of-speech

relation- 1
  type:          temporal-before
  from:          time-of-speech
  to:            time-of(clause-1)

```

Figure 6. TMR for the sentences in (10).

The frame *attitude-1* in the TMR indicates that the speaker has a fairly strong (value .8-1.0) deontic attitude toward *clause-1*, which says that the hearer goes somewhere unspecified. The syntactic structures for the English, Japanese, and Russian sentences in (10) are shown in Figure 7. These examples illustrate a constructional divergence in that the syntactic structures used by the individual languages to express the same meaning are radically different.

English Syntactic Structure

PREDICATE	had better
SUBJECT	you
COMPLEMENT	
PREDICATE	go
SUBJECT	you

Japanese Syntactic Structure

PREDICATE	ii (good)
SUBJECT	hoo (alternative)
REL-CLAUSE	
PREDICATE	itta
SUBJECT	pronoun

Russian Syntactic Structure

PREDICATE	stoit' (cost)
SUBJECT	tebe (you-dative)
COMPLEMENT	
PREDICATE	pojti (go)
SUBJECT	tebe

Figure 7. Syntactic Structures for the Sentences in (10).

---

(10) a. You'd better go.

b. Itta hoo ga ii.  
go-PAST alternative SUBJ good  
“The alternative that you went is good.”

c. Tebe stoit pojti.  
you-DATIVE cost-IMPERSONAL go-INFITIVE  
“To you costs to go.”

The lexicon entries in Figures 8, 9 and 10 show how the different syntactic structures are mapped onto the same TMR. We have indexed them by their most salient lexical item. The *syn-struc* fields characterize the language-specific realization of the construction.

For example, the *syn-struc* field of the English example says that this construction is headed by a verb *had* occurring with the adverb *better*, which takes a noun phrase as a subject and an infinitival clause as a complement.

The Japanese *syn-struc* field for *hoo* says that this construction is headed by an adjective such as *ii* or *tanosii* which is predicated of the noun *hoo* which is, in turn, modified by a relative clause in the past tense.

---

**HAD-BETTER**

```
Syn-Struc: predicate: [0]
           tense: past
           adverb: better
           subject: [1]
           complement: [2]
             subject: [1]
           inflection: infinitive

Sem:      clause   [3]
           head: meaning-of ([2])
           attitude [4]
             type: deontic
             value: 0.8-1.0
             scope: [3]
             attributed-to: *speaker*
             time: time-of-speech
           relation [5]
             type: temporal-before
             from: time-of-speech
             to:   time-of([3])
```

Figure 8. Construction Lexicon Entry for Expressing Deontic Modality in English.

---

In the Russian example (Figure 10), we are taking the verb *stoit'* to show impersonal agreement typical with the non-nominative subject *tebe*. Other analyses are possible. Again, it is important to note that the words *had*, *hoo*, and *stoit'* have other lexical entries corresponding to their other senses and uses. There are also many other construction lexicon entries corresponding to different constructions that also express deontic modality.

The *sem* fields in Figures 8, 9 and 10 contain TMR templates, which will give rise the TMR shown in Figure 6 when filled in. The English *sem* field indicates that the meaning of the complement of *had better* will become the main clause of the TMR. Similarly, the meaning of the complement of *stoit'* will become the main clause of the TMR. The Japanese entry for *hoo* indicates that the relative clause attached to *hoo* will supply the main propositional content in the TMR. In all three *sem* fields there is an additional component of meaning that says that the proposition expresses a high positive level of the speaker's deontic attitude toward the content of the proposition. The coindexings between the *syn-struc* and *sem* zones of these lexical entries will result in the same TMR being built even though the syntactic structures of the constructions are markedly different.

The above examples certainly cannot take the place of a full theoretical specification of the use of construction and should be viewed as a set of pre-theoretical intuitive considerations on the basis of which exploratory system development will occur. Armed

---

HOO

```
Syn-Struc: predicate: [1]
           root: (OR ii, tanosii. etc.)
           subject: [0]
           relative-clause: [2]
           tense: past

Sem:      clause [4]
           head: meaning-of ([2])
           attitude [5]
           type: deontic
           value: 0.8-1.0
           scope: [4]
           attributed-to: *speaker*
           time: time-of-speech
           relation: [6]
           type: temporal-before
           from: time-of-speech
           time-of ([4])
```

Figure 9. Construction Lexicon Entry for Expressing Deontic Modality in Japanese.

---

with the results produced by such an exploratory prototype we will, in our future work, proceed to formulating a more strict statement about treatment of constructions in a multi-lingual environment.

## 5 The Passive: Principled or Conventional?

In suggesting that vast numbers of constructions should be represented as entries in a construction lexicon, we are not recommending that the principled and rule-based aspects of language be ignored. In fact, our model of MT explicitly allows us to represent both compositional and conventional aspects of each construction. We will use the English passive construction to illustrate the interaction of the compositional and the conventional.

In reaction to older rule-based theories of syntax, proponents of modern principle-based theories have implied that the many constructions are figments of our imaginations. That is, the constellation of syntactic structures that make up a construction are not a unified phenomenon, but an accidental co-occurrence of independent phenomena that are each predicted by general principles. Some of the principle-based phenomena involved in the English passive are listed below. (See, for example, Levin (1988), Bresnan and Kanerva (1989) and Marantz, 1984.) We have attempted to present them in neutral way that is applicable to a number of different syntactic theories. They include, among other things:

---

STOIT'

```
Syn-Struc: predicate: [0] stoit'
           subject: [1]
           case: dative
           complement: [2]
               subject: [1]
               inflection: infinitive

Sem:      clause [3]
           head: meaning-of ([2])
attitude [4]
           type: deontic
           value: 0.8-1.0
           scope: [3]
           attributed-to: *speaker*
           time: time-of-speech
relation [5]
           type: temporal-before
           from: time-of-speech
           to: time-of ([3])
```

Figure 10. Construction Lexicon Entry for Expressing Deontic Modality in Russian.

---

- Morphemes, such as the passive participle morpheme, that unlink an agent or external argument from its syntactic position are common cross-linguistically. This is predicted by theories of the interaction of morphology and syntax or by theories of grammatical relations. It leaves the subject position open so that it can be filled by something else.
- Some principle of grammar determines that a direct object can become a subject when the agent or external argument has been unlinked (and there is not a locative or expletive element in subject position). This allows the active verb's direct object to correspond to the subject of the passive verb.
- Because the English past participle is not tensed, it must occur with a tensed auxiliary verb when it is in a main clause or any other environment that requires a tensed verb.

The following examples show that these are in fact three separate components to the passive, each of which can occur independently given the right circumstances. Example (11) illustrates passives without *be* in environments that do not require a tensed verb or that include another tensed verb. Example (12) illustrates unlinking of the agent argument without promotion of object to subject when a locative or expletive element is in subject position. Example (13) illustrates promotion of a direct object without passive in other constructions that involve unlinking of an agent or external argument.

- (11) a. *Admired by everyone*, she was sure to win the election.
- b. They got *arrested by the police*.
- c. We had them *arrested by the police*.
- (12) a. In this spot, well toward the center and front of the vast herd, appeared about to be enacted *a battle between a monarch and his latest rival for supremacy*. (Zane Grey)
- b. The wall paper was discoloured with age; it was dark grey, and there could be vaguely seen on it *garlands of brown leaves*. (W. Somerset Maugham)
- c. Here, in the stone wall, had been wonderfully carved by wind or washed by water *several deep caves* above the level of the terrace. (Zane Grey)
- d. Nowhere could be gotten *a better idea of its age* than in this gigantic silent tomb. (Zane Grey)
- (13) a. The bread cut easily.
- b. The glass broke.

In spite of these general, independent principles, there are strong reasons to view the passive as a unified construction for the purpose of machine translation. Although the components of the passive are each independently motivated, when they co-occur, they take on a range of meanings and functions that are not present when the components of the passive occur independently. The presence of the construction as a whole might, for example, signal certain interpretations of discourse focus or tense and aspect. These interpretations are neither inherent in nor unique to the passive construction, and may in fact require translation into different constructions in different target languages. Therefore, it is important to recognize the co-occurrence of the independent components so that specific meanings can be associated with the construction as a whole. In interlingual MT, those meanings should then be represented in the interlingua text in a way that is independent of the syntax of the English passive.

Processing of the passive construction can involve both a construction-based approach and a compositional syntactic analysis based on principles of syntactic theory. After a sentence has been parsed using compositional, theoretically motivated syntactic rules, the special co-occurrence of the independent components of the passive will be recognized by a construction lexicon entry such as the one in Figure 11.

The `sem` field of this entry indicates that the subject of the passive sentence is more salient than the oblique agent argument. It can also contain information related to other microtheories such as those of tense and aspect. This entry is indexed by the lexical item `be`. Other uses of passive verbs without `be` will be covered in separate lexical entries.

---

BE

```
Syn-Struc: predicate: [0] be
           subject: [1]
           complement: [2]
             subject: [1]
             oblique: [3]
               preposition: by
             inflection: past-participle
           voice: passive

Sem:      clause [4]
           head: meaning-of ([2])
           attitude [4]
             type: saliency
             value: .4
             scope: meaning-of ([3])
             attributed-to: *speaker*
             time: time-of-speech
           attitude [4]
             type: saliency
             value: .6
             scope: meaning-of ([1])
             attributed-to: .*speaker*
             time: time-of-speech
```

Figure 11. Construction Lexicon Entry for Passive Verbs with *Be*.

---

## 6 Conclusion

This paper presented a novel view of the boundary between the generalizable and the idiosyncratic in MT lexicons. We argue that the domain of the idiosyncratic should, in fact, be broader than in most current approaches. While at present most MT systems involve phrasal lexicons, these typically contain terminology from a particular field. In order to facilitate naturalness of translation, specifically, to carry the level of “conventionality” of meaning expression across languages, it becomes necessary to use the concept of a construction, a (possibly, discontiguous and productive) phrase whose meaning it is often impossible to derive solely based on the meanings of its components. It is also necessary to identify a construction in order to be able to select the most appropriate conventional way of expressing a meaning from among the available ways.

We discussed constructions in terms of the phenomenon of MT divergences. We have then shown how to incorporate the treatment of constructions into a standard interlingual MT environment, without losing syntactic or semantic generality of this approach. We claim also that treatment of constructions is both essential and attainable for the other major rule-based MT paradigm, the transfer approach.

## References

- [1] Bresnan, J., *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA, 1982.
- [2] Bresnan, J. and J. Kanerva, "Locative Inversion in Chichewa: A Case Study of Factorization in Grammar", *Linguistic Inquiry*, Vol. 20, 1989, 1-50.
- [3] Dorr, B., "Classification of Machine Translation Divergences and a Proposed Solution", *Computational Linguistics*, 1992.
- [4] Fillmore, C., P. Kay and M.C. O'Connor, "Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*", *Language*, 64, 1988, 501-38.
- [5] Fillmore, C. and P. Kay, *Linguistics X20: Construction Grammar Coursework*, Chapters 1-11. Unpublished lecture notes. University of California at Berkeley.
- [6] Levin, L.S., *Operations on Lexical Forms*, Garland, New York and London, 1988.
- [7] Marantz, A., *On the Nature of Grammatical Relations*, MIT Press, Cambridge, MA, 1984.
- [8] Matsumoto, Y. 1992. *On the Wordhood of Japanese Complex Predicates*. Ph.D. dissertation. Stanford University.
- [9] Meyer, I., B. Onyshkevych and L. Carlson, "Lexicographic Principles and Design for Knowledge-Based Machine Translation", CMU-CMT Technical Report 90-118, Center for Machine Translation, Carnegie Mellon University, 1990.
- [10] Nirenburg, S. and L. Levin, "Syntax-Driven and Ontology-Driven Lexical Semantics", in J. Pustejovsky and S. Bergler, (eds.), *Lexical Semantics and Knowledge Representation*, Springer-Verlag, 1992, 5-20.

# Dependency-Based Grammatical Information in the Lexicon

Petr Sgall  
Charles University, Prague  
*e-mail:* sgall@cspuk11.bitnet

## Abstract

Dependency grammar offers a suitable basis for including syntactic information into lexical entries, since a crucial point of dependency is the valency of every lexical unit. The valency (or case) frames, or theta grids, are to specify the possible (optional or obligatory) slots occupied by words which accompany a given head. Such a specification may also apply to adverbial and other "free" complementations. With such an approach, the word is understood as one of the central units of language. The class of (basically disambiguated) syntactic representations of sentences can then be specified with using just a small number of general principles to describe the core of grammar.

Don Walker's rich and fruitful work devoted to all aspects of the organization of lexical data invites us also to reconsider the task of specifying an appropriate way the syntactic information can be included in lexical entries. It appears that dependency syntax offers a suitable basis for this task, since a crucial point of dependency is the valency of every autonomous (autosemantic) lexical unit. For a dependency-based grammar it is natural to enumerate with each lexical unit the kinds of complementations (arguments and adjuncts) which can or must accompany the given unit as their head.

Dependency syntax, systematically treated first by Tesnière (1959) and briefly characterized in its main properties by Hays (1964), is to a large extent taken into account also in several theoretical approaches which overtly are based on another approach to syntax, namely on constituency. In fact, a treatment of valency can be found in Fillmore's Case Grammar, and certain important aspects of syntactic dependency are present in Bresnan's f-structure, in Relational Grammar, or in Chomsky's theta grids. Even earlier, such terms as 'head' and 'modifier', 'noun phrase', 'adjective phrase', etc., started to be used in transformational description, pointing to a similar approach; furthermore, Chomsky's X-bar theory has much in common with a dependency account of the basic syntactic relations.

With an approach using valency frames or grids, the word is understood as one of the central units of the language system. This point of view was reflected quite clearly in classical European structural linguistics, and during the development of modern theoretical linguistics the significance of the lexicon has steadily grown. It is now widely accepted that valency (or case) frames, or theta grids, are to specify the possible (optional or obligatory) slots occupied by words which accompany a given head. Such a specification may apply not only to 'arguments' or 'participants' (Tesnière's *actants*),

but also to adverbial complementations (his circonstants, including also the types of free adjuncts of a noun). These can be specified by means of a list, common to all verbs (or nouns, adjectives, etc.) and determined by grammar; this list can be activated whenever a lexical item is made use of by a generative procedure, by a parser, by a unification statement, etc., in a similar way as the data from the lexical entry are activated. A detailed discussion of a framework based on such considerations, the Pragian functional generative description, was presented in Sgall, Hajičová and Panevová (1986), where also possibilities of a classification of the complementations on the basis of operational criteria were discussed.

It is thus possible to work with a dichotomy, using the individual lexical entries (frames, grids) to enumerate the types of arguments, and, on the other hand, lists of adjunct types associated to the whole word classes.<sup>1</sup> Thus, from a certain viewpoint, with dependency syntax even a counterpart of Chomsky's X-bar theory can be basically placed in the lexicon. However, it is not necessary to use only a dichotomy; word classes and groups of different size can be specified e.g. by means of indices attached to the symbols for lexical meanings in the individual lexical entries and used to indicate the limitations of certain valency classes. Thus the class of verbs can be cross-classified into those having specific groups of arguments (each of which is characterized as optional or obligatory), those requiring obligatory adjuncts (see below), having specific subcategorization properties, and so on.

In other words, the boundary between lexicon (individual entries) and grammar (with a syntactic classification of words) is not immediately given and is more complex than it might seem.

Another, often neglected, aspect of this complexity of the boundary concerns the status of function morphemes, which, in the outer shape of sentences, are divided between words (prepositions, subordinating conjunctions, articles, etc.) and sublexical morphemes (affixes, endings, alternations). Languages widely differ in using affixes, endings or function words as the means of expression of grammatical values and functions. From the viewpoint of a general linguistic theory it is not immediately theoretically relevant whether, e.g., cases (roles), tenses, adverbial relations are expressed more often by prepositions, auxiliary verbs and subordinating conjunctions (and word order), as is the case in English or French, by inflectional endings (and prepositions), as is the case in most Slavonic languages, Latin, etc., or by affixes, as in Finnish, Hungarian, and so on. While these differences concern the surface (or morphemics) of individual languages, it is relevant for the general theory that function morphemes, whether they have the surface shape of words or not (which in some cases is rather unclear, cf. the French clitics—or prefixes—in, e.g., *je t'en donnerai*), are not syntactically freely combinable;<sup>2</sup>

---

<sup>1</sup>We assume that every verb can be accompanied by any free complementation, i.e. that there are no grammatical constraints restricting e.g. a directional adverbial just to verbs of motion (since there are also such sentences as: *The carpet was lying from the window to the door*). If it is difficult to find examples with final clauses or adverbials of means accompanying verbs the content of which is not connected with human intention, such difficulties are due to semantic relationships, rather than to grammatical constraints. This also concerns e.g. the boundary line between strict subcategorization and semantic selection restrictions. Shooting ashtrays, plants growing for some goal to be achieved, or cannibalism are not to be excluded by grammatical means.

<sup>2</sup>Let us add that it is impossible for a syntactic account to understand all lexical units as determined immediately by the outer shape of sentences. In any case, such word forms as those within which an article

a preposition or an article is always combined with a noun, an auxiliary verb or a conjunction accompanies a verb, etc. (with exceptions which have to be described as such). Thus, function morphemes can be adequately described by indices of node labels, or by parts of complex symbols. In fact, also in Chomsky's Principles and Parameters Theory, the difference between function words and sublexical grammatical morphemes is not understood to be crucial; both these classes are treated in a similar way. The main difference is that, while Chomsky works with an analytical metalanguage, comprising such nodes as Aux, Compl, Agr, Infl, the metalanguage we assume to be preferable is agglutinating: the correlates of function words have the shape of affixes, of indices in the sentence representations. An advantage of the latter approach is that the structure trees can be more simple, not containing more nodes than necessary.

It also is important to bear in mind that the opposition of arguments and adjuncts concerns the kinds of complementation as such; on the other hand, the difference between obligatory and optional complementations characterizes the relationship of a complementation to its head (as a lexically specific unit), rather than the kind of complementation itself. It is possible to find arguments which with certain heads are not obligatory (e.g., the Addressee with *to read*), and also adjuncts which with a certain head are obligatory, e.g., *to behave somehow, to arrive at a place, to last some time*. It is not important that in some cases such obligatory complementations can be deleted in the outer shape of the sentence (especially if they are determined by the context).

In an approach based on the principles briefly outlined above, a lexical entry may contain the following parts:

- (a) A representation of the lexical unit itself, i.e., of its lexical meaning. In case of ambiguity there are several representations (i.e., several lexical entries), whereas in case of vagueness we have a single meaning (vagueness, as a property of meaning, is - partially - resolved during semantic interpretation, by inferencing based on contextual and other knowledge).
- (b) A specification of the values of relevant grammatical categories, i.e., of grammemes belonging to the given word class (e.g., number and definiteness with nouns, or tense, aspect, different kinds of modalities, etc., with verbs, degrees of comparison with adjectives). Restrictions on the combinations of these values are listed for every word class as a whole, only exceptions have to be registered in individual lexical entries (or, by means of indices, with smaller groups).
- (c) The valency frame of the given lexical unit, the basis of which is the list of its possible complementations (Actor, Addressee, Objective, Effect, Origin, Locative, Instrument, Manner, Cause, etc.). In the frame, arguments and obligatory adjuncts are indicated. Since they may be either deletable (as Directional-2 with *to arrive*) or not deletable (as Objective with *to create*), it is denoted by a specific index

---

is joined to a preposition (G. *am, im, beim, ans*, etc., or Fr. *du, des, au, aux*) make it necessary to work with a "hidden" boundary, i.e., with "underlying" words divided from each other in a way different from the "overt" word shapes; cf. also such combinations of auxiliaries with other items as E. *won't, ain't*, Czech *spals-li* 'if you slept', *abys* 'for you to', and so on. The question then is what criteria are to be used to establish this hidden boundary. The distinction between autosemantic lexical units and function morphemes, which is an intrinsic ingredient of any linguistic theory distinguishing between lexicon and grammar, may be useful as such a criterion.

whether a complementation is deletable. Also the optional or obligatory function of an item as controller is specified here (e.g., Actor is an obligatory controller with *to try*, an optional one with *to decide*; Addressee is an optional controller with *to advise*, *to forbid*). Furthermore, indices of the individual complementations characterize them as being able to occupy certain specific positions in the clause (e.g., that of subject, or of a *wh*-element) or to appear as barriers for movement.

- (d) Subcategorization conditions determined with the individual kinds of complementations in the frames, e.g., the Objective of a verb possibly having (or not) the shape of a noun group, of a verb clause, etc.

In some cases different lexical entries share their lexical part proper, i.e., the lexical unit. Such lexical entries differ only in their frames, which provide different starting points for semantic interpretation. This concerns such verbs as *to swarm* (either with an obligatory Means as in *The garden swarms with bees*, or without it as in *Bees swarm in the garden*) or *to load* (either with an obligatory, though deletable Means as in *They loaded the truck with hay*, or with an obligatory Directional-2 as in *They loaded hay on the truck*).

If lexical entries of the shape outlined above are used, then the class of representations of sentences can be specified with using just a small number of general principles to describe the core of grammar. To illustrate this point, we present an outline of the main points of a procedure generating sentence representations (a detailed elaboration can be found in Petkevič, in press):

1. To generate a node *n* means
  - (a) to create the node *n* either as the root of a representation, or as a node that is dependent on another one and is placed to the right of all its sister nodes, and also
  - (b) to choose *n*'s lexical value and the values of its grammatemes; this is to be done taking into account the subcategorization conditions of the mother node and the restrictions on the combinations of grammatemes (specified in the lexical entry of the head or in the data concerning the respective word class); the technique used to realize these conditions and restrictions is unification; e.g., if *n* is the Objective of a verb that subcategorizes its Objective as a verb, then the lexical unit in the label of *n* has to display an index identifying its word class as verb;
  - (c) if *n* is a root, the lexical part of its label is a verb, and its grammatemes determine it as a finite verb form of the main clause; *n* is then specified either as CB (contextually bound, i.e., as belonging to the topic) or as NB (non-bound, i.e., as belonging to the focus).
2. If the symbol of a complementation (inner participant or adjunct) is present in the frame of the node *n*, then it is possible to generate either a left or a right daughter of *n*:
  - (a) in case a left daughter is being generated, a CB marker and a complementation value chosen from the frame of *n* is assigned to it;

- (b) if a right daughter is being generated, it gets a NB marker (which can be understood as primary, i.e., as the absence of a marker, similar to the case with all the primary values of the grammateemes) and a complementation value chosen from the left end of the frame of  $n$ .

Note: If the chosen complementation is an argument, it is deleted in the frame of  $n$  (as having been saturated). Choosing a complementation “from the left end” means that optional complementations can be skipped; the chosen and also the skipped complementations are deleted in the frame of  $n$ ; if the last one present there has been deleted, no more daughter nodes can be generated in this step, and point (3) below is carried out. The order of the complementations in the frame is determined by their systemic ordering.<sup>3</sup>

3. If no complementation is present in the frame of  $n$ , then the procedure goes back to the mother node of  $n$ , which now is to be considered as node  $n$ ; if no mother node is present, the procedure is finished.
4. Only representations containing a focus are understood as representations of sentences, more exactly, only those whose branch going from the root to the rightmost daughter of the rightmost daughter of ... of the root includes a NB node.

As Petkević (in press) has shown, a declarative specification of the class of the underlying representations of sentences can follow similar lines, using unification, if the latter concept is so complemented as to allow for checking the order of the items depending on a single head and for distinguishing between saturated and non-saturated items (cf. the deletion of a saturated inner participant, mentioned in the Note above, which ensures that an inner participant occurs at most once in a clause; this restriction does not concern free, adverbial complementations, cf. the three Temporal adverbials in *Yesterday she came late to the office in the morning*).

This specification of the underlying representations covers only the core of sentence syntax. It has to be completed in several respects, especially in what concerns coordinated structures (corresponding to a third dimension of the network), the positions of such syntactically specific items as the operator of negation and other focalizers (*only, even, also, etc.*), and the marked cases of word order (see fn. 3).

It thus appears to be possible, if dependency syntax is used together with a “free” word order in the syntactic representations and if a large portion of grammatical information is included in lexical entries, to describe the core of syntax in a relatively economical way. The general character of the principles of this description makes it into a useful alternative

---

<sup>3</sup>The systemic ordering, discussed in Sgall et al. (1986, Chapter 3) determines the prototypical order of complementations; deviations of the word order from systemic ordering are due (i) to the CB items being placed to the left from the focus, (ii) to grammatical word order rules, such as those concerning the positions of the verb, of the adjective before (or after) its head noun, of the clitics, etc., and (iii) to a marked placement of the focus proper (bearing the intonation center of the sentence), e.g. in *Yesterday SHE came late to the office*. In English, the systemic ordering of some of the main complementations is as follows: Actor – Addressee – Objective – Origin (Source) – Effect (Result) – Manner – Directional.1 – Means – Directional.2 – Locative. Thus the prototypical order is present e.g. in *He came by car to a lake*. On the other hand, in *He came to a lake by car* the Directional.2, being CB, stands more to the left than what would correspond to its position in systemic ordering.

to Chomsky's Universal Grammar; a highly natural account of innate properties allowing for the acquisition of language as embedded in context may be gained in this way.

## References

- [1] Hays, D. G., "Dependency theory: A formalism and some observations", *Language* 40, 1964, 511-525.
- [2] Petkevič, V., *Underlying structure of sentence based on dependency*, Charles University, Prague, in press.
- [3] Sgall, P., E. Hajičová, and J. Panevová, *The meaning of the sentence in its semantic and pragmatic aspects*, ed. by J. Mey, Reidel, Dordrecht—Academia, Prague, 1986.
- [4] Tesnière, L., *Eléments de syntaxe structurale*, Klincksieck, Paris, 1959.

# Semantics in the brain's lexicon — Some preliminary remarks on its epistemology

Helmut Schnelle  
Sprachwissenschaftliches Institut  
Ruhr-Universität Bochum  
*e-mail:* schnelle@linguistics.ruhr-uni-bochum.de

## **Abstract**

A discussion of basic properties of a new, neurobiologically motivated linguistics based formally on dynamical systems theorizing rather than on combinatorial symbolic structuring.

## **1 Mental Models or Brain Models?**

Much of current theoretical work on lexica claims to be about mental lexica, i.e. lexica in the mind or even lexica formalized in the framework of a computational theory of mind. The guide-line for modern conceptions of mental entities (e.g. "mental word") is the same as that of classical rational conceptualism or formal logic: they are recursive compositions of conceptual elements. A recent variety of this position has been presented by Jackendoff (1987). He seems to imply, like the majority of theoretical linguists, that this kind of analysis is a necessary step towards a more complete description of the human mind/brain: higher cognition (like language) developed in terms of the computational theory of mind is a prerequisite of higher cognition understood in terms of the computational theory of brain.

This perspective is increasingly influential also in lexicography. Instead of seeing lexica as lists of information about isolated words, they are rather understood as computational systems in which lexical units are structured entities generated by the compositional principles of the lexical component. A more direct approach to the analysis of how words and sentences are in the brain is possible but practically nonexistent, due to a general agreement among linguists that we lack neuroanatomic and neurophysiological details. Thus, brain-style theorizing is considered to be premature. (Cf. the discussion in Schnelle, 1994).

This conclusion is not cogent at all. In the case of mental linguistic theorizing, theoreticians did not wait for a sufficient clarification of psychological facts. Rather, Chomsky and other linguists proposed principles of analysis, description and notation which, subsequently, contributed to determine what cognitive science is about. Thus, it was not that theorizing emerged from empirically established psychological facts; rather, theorizing contributed to the definition of hypotheses and frameworks for empirical research. Correspondingly, we need theoretical principles of analysis, description and

notation that imply empirical hypotheses to be tested by empirical neuropsychological and neurobiological research. These hypotheses should specify in particular how sounds, words and sentences could be conceived in the brain—given current knowledge of brain architecture and processing. They should be stated in formats that could be easily interpreted by the sciences involved.

What are the possible principles? It is not necessary to start at zero. Just as Chomsky could build on results of foundational research on formal systems (e.g. Post-calculi with their concept of a rewrite system and the more general theory of recursive algorithms) we may now build on

- dynamical systems models of the computational brain (e.g. Sejnowski and Churchland, 1989, Churchland, 1989, Churchland and Sejnowski, 1992, Kosslyn and Koenig, 1993),
- neuroanatomic and neurophysiological investigations into the localization of linguistic objects and processes in the brain (cf. Caplan, 1987, Part II and IV, Pulvermüller, 1992),
- neuropsychological research on partitioning of brain areas into cooperative functional units (cf. Damasio, 1989, Edelman, 1989, Edelman, 1992, Kosslyn and Koenig, 1993, etc.),
- structural linguistic considerations, with reference to aphasia research (cf. Caplan, 1987, Part III, Deane, 1992).

As essential features of theoretical and empirical approaches to brain representation we thus need the following:

1. a basic representational *formalism* (a formal mathematical representation of the dynamics, either in terms of symbolic computation or in terms of causal laws),
2. its interpretation in terms of *neuroanatomic localizations* in the brain (either in neuroanatomic terms of gyri, lobes, Brodmann areas, etc. or in terms of brain coordinates as used in neurological imaging techniques),
3. its interpretation in terms of *neuropsychological functional unit localizations* in the brain, and
4. its interpretation in terms of descriptive *labels* (understood as linguistic features, categories and structure representations).

There is a particular difference between the symbolic and the causal approach which concerns the role of the symbols. In the symbolic approach symbols and symbolic representations would not merely serve as interpretative labels on level 4, but are themselves elements in terms of which the dynamics of computation is specified and actually controlled. In the approach aiming for a causal analysis, the features and categories are merely labels that contribute to the understanding of the analyst by marking the functional architecture of the network (similar to the labels on blue-prints and logical

designs of electronic networks); they do not actively control the processing (say by being represented by bits controlling the processes themselves as do the symbols in a symbolic algorithm).

However, theorizing in terms of dynamical systems is so unfamiliar to most linguists that it will be necessary to show in more detail how it differs from the more customary approaches. We shall do this by discussing some basics of externalization of language as proposed by Quine as well as the basics of mental representation as outlined by Jackendoff which exemplify with epistemological sophistication how linguistic theorizing can be integrated into the general methods of scientific theorizing.

## 2 The Scientific Status of Linguistic Theorizing

We follow Pollard and Sag's (in press) suggestion on theoretical mathematical analysis in a formally analyzed discipline:

In any mathematical theory about a domain, the phenomena of interest are modelled by mathematical structures, certain aspects of which are conventionally understood as corresponding to observables of the domain. The theory itself does not directly talk about the empirical phenomena: instead, it talks about, or is interpreted by, the modelling structures. Thus the predictive power of the theory arises from the conventional correspondence between the model and the empirical domain.  
(Pollard and Sag, in press, Introduction)

Pollard and Sag continue by referring to examples from physics:

... in one kind of standard model of celestial mechanics, the positions and velocities of bodies, subject to mutual gravitation are represented by vectors in a higher dimensional Euclidean space ('phase space'), the masses of the bodies by positive real numbers, and their motions by paths along certain smooth vectorfields ('flows') on the space. Of course, such a model is not the same thing as what it models (e.g. the solar system), but certain formal properties of such a model may represent aspects of the solar system of interest to the physicist. In a formal theory of such a model, the underlying logic is just a standard first order language (e.g. the Zermelo-Fraenkel set theory) and the axioms are certain systems of differential equations (e.g. Hamiltonian systems) that the flows are required to satisfy. An observed motion of the solar system is then predicted by the theory insofar as it agrees—under the conventional correspondence—with an admissible flow (i.e. one that satisfies the equations). (ibid.)

Pollard and Sag represent these facts by a figure in which the range of phenomena (possible motions of  $n$ -body systems) is modelled by a mathematical schema (Hamiltonian vectorfields) which is taken to be a model-theoretic interpretation of the formal theory underlying the mathematical model. The mathematical schema has an important function in mediating between the formal theory and the phenomena. (Cf. also Schnelle, 1991, Ch. 3 and 8.)

A natural continuation of this exemplification would be to take the activities of neurons in the brain (in particular in those brain areas relevant for speech and understanding)

to be the observables. An actual brain state is, then, determined by the actual states of activities of the neurons just as the state of a mechanical system is determined by the actual states of the bodily positions and velocities. The dynamics of a mechanical system is determined by its differential equations; correspondingly, the equations defining the reactivities of the neurons define how changes of neuron activities depend on the activities of the neurons they are connected to. The set of equations for all neurons determines the “flows” in the corresponding phase space. The specific parameters of the equations are embodied in the system by a process of the acquisition of the dynamics. Or, as de Saussure said, they are the traces which represent “*la langue*” in the speaker’s brain (de Saussure, 1962, Intro., chap. III, §2, chap. VI, §1). In this Saussurean interpretation, a language (*la langue*) is the network of dynamic equations which determine the linguistically relevant activities in the brains of the speakers of the language. This holds in particular for the lexicon and for lexical semantics. It comes as a surprise that Pollard and Sag, after having provided their physicalistic introduction, do not continue as indicated but propose non-dynamic entities as observables, namely the sound events or sound types as utterance-products or forms of utterance-products. The proper physical paradigm for their discussion would have been statics rather than dynamics. Correspondingly, the mathematical models assumed by Pollard and Sag are not positions and trajectories in a linguistic phase space but static patterns: sorted feature structures. But Pollard and Sag’s proposal is rather standard in linguistics in claiming that recursively defined sets of static structures underly “somehow” the processes occurring in human users.

Let us once again return to the physicalistic paradigm and note that there is another way to interpret it. It is the famous metaphor from Quine (1953) which unfortunately was not taken up in linguistics. Quine proposed the following idea: “The totality of our so-called knowledge or beliefs, . . . is like a field of force whose boundary conditions are experience” (Quine, 1953, p. 42). Quine elaborates on this suggestion and also applies it to the nerve net as the field of force (Quine, 1960, p. 74, p. 94, and elsewhere).

But considering logic his main field of research Quine does not “dig deep into the brain” and prefers to *externalize the epistemological issues*. He replaces the nerve net interpretation of the metaphoric field of force by a “web of beliefs” represented by inferentially related statements or sentences. Thus, instead of relating sensory stimuli to brain states, he externalizes the model by relating its components to sentences, such that the sentences form the nodes of a network in which the connectivities represent the logical and semantic interrelations. In this way Quine maps connectivities inside the brain onto relations defined over publicly observable units (i.e., sentence utterance patterns). Recently, he concedes that “neurology is opening strange new vistas into what goes on between stimulation and perception . . . [and] the passage from perception to expectation, generalization, and systematization. . . . Within this baffling tangle of relations between our sensory stimulation and our scientific theory of the world, there is a segment that we can gratefully separate out and clarify without pursuing neurology [etc.] . . . its essentials can be schematized by little more than logical analysis” (Quine, 1990, p. 1/2) Quine provides the bridge from empirical facts to logical analysis and epistemology by a *twofold externalization*—external utterance and socialization (i.e., binding to a community of speakers having access to the publicly observable sentences). This position is now standard among logicians and epistemologists. According to this

perspective, “the totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only at the edges” (Quine, 1953, p. 42). Thus the internal neuronal “field of force” is replaced by an external “man-made fabric”, a formalized theory, which consists of the set of interdependent statements on which the community agrees. Instead of describing and using reference to internal language representations, as Chomsky does, Quine describes and uses regimented forms of external language. His proposals are very useful in analysing the lexicographer’s view of semantics, as I have explained elsewhere (Schnelle, 1991, Ch. 11). Nevertheless, the mainstream of linguistic research accepts the task of linguists to be that of determining the mental representations which constitute the language in the minds of speakers. A clear characterization of what is involved has been given by Jackendoff. We shall now turn to the discussion of his proposals before coming to the more biologically minded proposals made by the Churchlands and others.

### 3 Mental Representations According to Jackendoff

Jackendoff argues that the theory of mind should be a computational theory. This involves the following assumptions:

1. We should distinguish between the brain as a physical or biological entity and mental entities.
2. In contrast to certain classical varieties of mentalism interpreting mind as phenomenologically given and confronting it directly with the body, we should consider computational theories, in which that relation is mediated by the notion of a computational mind (Jackendoff, 1987, p. 20).
3. The computational mind should be determined by “the information processed by the brain and the computational processes the brain performs on this information” (*ibid.*, p. 15).
4. The “information” processed is further required to have a “form” (p. 38) which, in “the structural approach to the computational mind” (*ibid.*) is defined on the basis of “categories, distinctions and relations” that “must be represented in mental information structures in order to account for human behavior and experience” (*ibid.*). This form of the information processed is the mental representation “contained” in the brain (p. 29). Thus the brain is primarily seen as a container of mental representations and secondarily as a processor whose processes are controlled by these mental representations (p. 44).
5. Mental representations form a set of “structured repertoires of distinctions that can” not merely be determined formally by means of a combinatorial and recursive definition, but can “be encoded by the combinatorial organization of the computational mind” (p. 47). More precisely, the combinatorial organization is such that the “structured repertoire is built up from a finite set of primitive distinctions, plus a finite set of principles of combination that make it possible to build

primitives into larger information structures" (ibid.). The complete set of all information structures that can be built from primitives constitutes a "level of mental information structure" or a "level of representation" (ibid.). (In this last point, Jackendoff approaches the classical epistemological abstraction of formal logic: rational construction of mental entities modeled by symbolic combinations.)

Most linguists would of course take these assumptions as matters which do not need to be made explicit at all. The last point (5) is particularly typical for many approaches to the lexicon in which the systematic structure of a lexicon can only be made clear by describing the lexicon as a combinatorial *level of representation*. Each word is to be assigned a word structure built from the primitives (e.g., phonological, morphological, syntactic and semantic) appropriate for the description of lexemes, i.e., words in the lexicon. The construction must be properly represented by structural means (bracketings, trees, lattices, etc.).

Contrary to the mainstream in linguistics, I consider these assumptions to be very misleading when we are concerned with finding out how words are represented in the brain. Their problematic status is revealed by a closer examination of Jackendoff's examples of non-linguistic and linguistic representations. Thus, after having introduced the notion of a level of representation, he tries to illustrate it by the retinal array: "Here the primitive distinctions might be (roughly) position, light, intensity, and color. The principles of combination might be (1) identifying a position as encoding a particular intensity and colour at a given time; (2) identifying the spatial relation of two adjacent positions (above, below, left of, and so on). From these primitives and principles of combination the state of the entire retinal field at a given time can be encoded" (p. 48). This is certainly a possible feature representation (1) together with a topological organization (2). But there is no (recursive) principle of combination as seems to be required by Jackendoff's general definition of the notion level.

In another passage, Jackendoff tries to correlate this schema of retinal representation with typical phonological representation. The latter is not, according to Jackendoff

completely specified in terms of a set of primitives and principles of combination. The primitives are the distinctive features and the notions of phonological segment, syllable and word. The principles of combination are (1) the simultaneous combination of a compatible set of distinctive features into a matrix for a single phonological segment; (2) the concatenation of segments; and (3) the bracketing of concatenations of segments into syllables and of syllables into words. Out of these primitives and combinations can be built both underlying and surface phonological representations of any utterance in any language. (p. 63)

(1) may be easily correlated with the formal specification of (1) in the retinal case but (2) and (3) have no such direct counterparts. Assume that we have a phonological "retina". In it the feature bundles (case (1)) should be represented by activity positions of some (neuronal) group. The features in a bundle stand in a relation of "simultaneous concurrency" (as suggested earlier by Jakobson). The topological organization (corresponding to (2) in the retinal case) should be given by "immediate sequential contiguity" for segments or features (also suggested by Jakobson). Having an array of features ordered by two topological relations does not yet provide a representation of units. We might

well postulate (neuronal?) entities whose activity would then represent the presence of combined units, such as phonemes, syllables and words. Here, a finite number of units will be introduced—corresponding to a finite list of learned items.

However, our system could not represent an *infinite* number of units, nor would it have a method to construct an unlimited number of units on the spot. Certainly, the system can *learn* more and more units; thus by *learning* the system can be *extended* practically without limit. But never does the system reach a stage in which it would have a “faculty of use” of an unlimited number of units, e.g., of potential syllables or words, *immediately available*.

This now is the essential difference with Jackendoff’s suggestion that the system has concatenation available: concatenation is a recursive (and thus unlimited and infinite) operation of use, whereas the topological relation of immediate contiguity merely defines a finite relation over a finite perceptual array (units for feature representation, syllable representation, or word representations, which stay active a sufficiently “long” time for the current morpheme representation). Correspondingly, there cannot be any unlimited, combinatorially determined bracketing of strings concatenated on the spot. Instead, we have additional units whose activity registers the current activity of a proper configuration of constituent units (the presence of a pattern which the current standard representation expresses by a bracketing). There is no generativity of use; at any time, the set of spontaneously available (and representable) language patterns is finite and limited. There is certainly generativity of learning; at any time, the range of learned and spontaneously available units can be extended. This holds for spoken language. The situation is different for consciously controlled operation on written text. There, the notion of strict generativity *is* applicable but it applies only to the taught skills of written language, in which *checking processes* are executed *with pencil and paper*. These skills do not grow naturally in first-language acquisition.

In view of the danger of misleading theorizing, we should *avoid*, in our theoretical metalanguage, *the use of rules or operations which can be executed over unlimited stores of symbol representations*; i.e., we should avoid our standard formalisms or data-type specifications. We should replace them by the descriptive means of dynamical theories. Let us now discuss some of their characteristics.

## 4 Features of Linguistic Structure and Processing as Features of Dynamical Systems

As in the case of symbolic logic and theories of symbolic formalisms and symbolic computations, our proposals should again be developed in the context of epistemology. But we should note that *epistemology can—and should—be naturalized, and thereby internalized*, as Quine suggests. (Cf. Quine, 1969, Ch. 3.) More and more proposals have been made recently. Details can be studied in P.M. Churchland (1989), P.S. Churchland and Sejnowski (1992), and others. P.M. Churchland shows “how a representational scheme can account, in a biologically realistic fashion, for a number of important features of motor control, sensory discrimination, and sensorimotor coordination.” But he continues to ask: “has it the resources to account for the so-called higher cognitive activities, as represented by *language use*, for example, and by our propositional knowl-

edge of the world in general?" (P.M. Churchland, 1989, p. 109). He believes that this is possible in the representational schema he proposes. "The basic idea is . . . that the brain represents various aspects of reality by a *position* in a suitable *state space* and the brain performs computations on such representations by means of general coordinate transformation . . ." (ibid. p. 78/79). In this context, Churchland suggests "a way of representing 'anglophone linguistic hyperspace' so that all grammatical sentences turn out to reside in a proprietary hypersurface within the more general hyperspace, with the logical relations between them reflected as spatial relations of some kind. I do not know how to do this, of course, but it holds out the possibility of an alternative to, or potential reduction of, the familiar Chomskyan picture" (ibid. p. 109). The dynamical systems terminology used in this passage may well illustrate to the linguist how foreign such physicalistic formalisms are compared to the formalisms he or she is used to. In any case, being able to cope with linguistic structure is accepted as a crucial challenge by brain analysts. (Cf. apart from the Churchlands, Sejnowski, Edelman, Kosslyn/Koenig.) Churchland's proposal has recently been criticized by Fodor and Lepore (1992). With respect to the general claim, they argue that the approach cannot be both an alternative to and a reduction of Chomsky's picture. "Qua *alternative* it would, by definition, be in competition with Chomsky's picture. A *reduction*, by contrast, would presumably be a story about the neural format in which what the native speaker knows about his language is coded in his brain. By definition, a reduction is not in competition with the theory that it reduces" (Fodor and Lepore, 1992, p. 201). I believe that this critique is not justified. The reduction of classical thermodynamics to statistical mechanics certainly changed some aspects of thermodynamics (cf. Arbib and Hesse, 1986, p. 64). Correspondingly, we would have to expect changes in linguistic conceptions also if reduction to brain processes were to be successful. Thus trying to provide a reduction of symbol representational linguistics to dynamical systems linguistics with neuron activities in the brain as the observables is not only a methodological alternative to the Chomskyan picture but also a *factual alternative to it which has the potential of fruitful but unforeseeable consequences*. Fodor and Lepore discuss at great length what can and what cannot be represented by the new approaches. I believe that it is rather premature to go into much detail here. The possible alternatives are not sufficiently detailed and vivid for precise discussion. We must first develop and test the new ways of representing linguistic processes in brain-style fashion. After all, language is in the brain and a proper understanding of the nerve nets in the brain *and* of their functional architecture *should* be able to show how language functions. Rather than applying much intelligence to the discussions of what can and what cannot be done in principle, we should work hard on descriptive techniques, whether connectionist, implementational or whatever. Thus, I shall not enter into further discussion of what can and cannot be done but rather suggest some descriptive avenues towards semantics in terms of nerve nets.

## 5 Formalisms for State Space Semantics?

The suggested formalisms are state space descriptions. What is a state space? Usually, this concept is presented in rather formal mathematical contexts. Intuitively, state space representations apply to a collection of objects which have observable properties. (Compare the basic discussions in Schnelle, 1991, Ch. 3 and 8.) As already discussed

in connection with the proposal by Pollard and Sag, the classical examples from physics are mechanical systems consisting of a set of particles with positions and velocities as observables. For perceptual psychology the object is the perceiving individual and the observables are taken to be visual, auditory and tactile properties of entities of perception. The possible values for the observables are real numbers for positions and velocities referring to space or the colors red, yellow, green, blue, ... as values for the color perception. Formally: If the system to be described has  $n$  objects and each object can be characterized by  $m$  observable properties, one introduces a formal "space" spanned by  $n \times m$  "axes" or "dimensions" and calls it the state space of the system. Just as a point in real space is determined by specifying a particular coordinate value for each of the three axes, a point in state space is determined by specifying a value for each of the  $m$  observables for each of the  $n$  objects in the system. A point in the state space represents one of the empirical possibilities of the system and the set of all points represents all empirical possibilities. Correspondingly, one also sees the state space as a *possibility space*, in particular in cases in which probability values are assigned to each point, resulting in a probability distribution over the space. Applying these concepts to the brain, we may say that a part of the brain comprising  $n$  neurons (such that each neuron can be characterized by one observable, its momentary activity) can be represented by an  $n$ -dimensional state space. Changes of the activities of the neurons are represented by a trajectory in the space, leading from an initial point to a final point. The changes are determined by the dynamic of the state space. In typical cases, the dynamic is determined by regularities or laws, which determine for each observable how its change of activity is determined by the activities of certain other units to which it is connected (the latter are its functional neighbors).

How are we going to relate these ideas to conceptualizations of linguistic structures and processes? The first model which I developed involved the following procedure: a rule system (e.g., constituent structure grammar) is translated into another representational format, the so-called dotted rule representation of an Earley parser. Occurrences of dotted rule units are then assigned to activity units ("neurons"). The set of units introduced is the set of observables spanning the state space. A proper analysis of a sentence is a sequence of changes of the observables which leads to a final state (attractor state) representing the configuration of those units of dotted rule symbols which represent the complete and successful parse output (cf. Schnelle, 1991, Ch. 7, Schnelle and Doust, 1992, Wilkens and Schnelle, 1990). Another approach is currently being studied by our research group. It interprets the nodes of a graph notation for a feature structure as the observables to be represented by neuronal groups. Each graph of a particular feature structure is a subgraph of a much larger array of potential alternative graph connectivities which gets activated just in case the corresponding symbol is present. Given a grammar which expresses its statements in terms of feature structures, our task would be to generate a complete graph which contained each particular feature structure as a subgraph. Or, seen from the other side, the complete graph is the superposition of all graphs representing single structured units. We would also be interested in assigning to the connectivities dynamic conditions in such a way that the activity of an input configuration for a feature structure (say the activation of the phonetic representation of a word) would trigger a process which generates the features structure which belongs to the word. If the input of a sentence is given, activity patterns representing the words

should be activated one after the other and patterns representing the more “abstract” or more “internal” syntactic process of expectations and satisfactions of grammatical category occurrences become activated in a proper pattern of a network of neuronal syntactic control units. The details of this translation of feature structure grammars into neuronal networks are rather intricate, but it seems that the translation is tractable.

Our research focuses in particular on the embodiment of the lexicon or on how a linguistically specified framework can be interpreted as a formal state space which can be mapped into a neuronal state space of the brain. Obviously the state space should have the property that the similarity between linguistic feature representations of different lexemes should somehow correspond to similarity relations concerning the localization and connectivities of the neuron patterns in the neuronal space of the brain.

To ensure the empirical breadth of our study, we are simultaneously working on the translation of information of ordinary dictionaries into HPSG feature structures (cf. Hoelter, 1993) Up to now, processing of several hundred entries of different types have shown how the entries of the Cobuild dictionary can be automatically translated into HPSG structures.

## 6 Conclusion

We need a new perspective for linguistic theorizing. It should bring linguistics closer to dynamical systems modelling, freeing it from the spell of logical and symbolic formalisms. The latter have been developed in view of pencil and paper operations (from Hilbert to Post, Turing et al.) which have been adapted to the task of constructing a symbolic-program-controlled universal machine, the computer (cf. Schnelle, 1987). Natural languages are connectivities in the brain which provide for highly parallel, efficient and flexible language use that is basically different from consciously controlled activities of rational argumentation. What we need are means for theorizing which are adapted to such systems.

## References

- [1] Arbib, M.A., Hesse, M.B., *The Construction of Reality*, Cambridge University Press, Cambridge, 1986.
- [2] Caplan, D., *Neurolinguistics and Linguistic Aphasiology*, Cambridge University Press, Cambridge, 1987.
- [3] Churchland, P.M., *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press, Cambridge, MA, 1989.
- [4] Churchland, P.S., Sejnowski, T.J., *The Computational Brain*, MIT Press, Cambridge, MA , 1992.
- [5] Damasio, A.R., “Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition”, *Cognition*, 33 , 1989, 25–62.

- [6] Deane, P.D., *Grammar in Mind and Brain: Explorations in Cognitive Syntax*, Mouton-de Gruyter, Berlin, New York, 1992.
- [7] Edelman, G.M., *The Remembered Present: A Biological Theory of Consciousness*, Basic Books, New York, 1989.
- [8] Edelman, G.M., *Bright Air—Brilliant Fire: On the Matter of Mind*, Penguin Press, London, 1992.
- [9] Fodor, J., Lepore, E., *Holism: A Shopper's Guide*, Blackwell, Oxford, 1992.
- [10] Hoelter, M., “The BLF analysis”, in ET-10/51 Group, “The Parsing of Cobuild Definitions and Mapping of the Output to Typed Feature Structures and Bochum Logical Form”, ET-10/51 Report 4, Commission of the European Communities, Luxembourg, and Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, 1993.
- [11] Jackendoff, R.S., *Consciousness and the Computational Mind*, MIT Press, Cambridge, MA , 1987.
- [12] Kosslyn, S.M., Koenig, O., *The Wet Mind: The New Cognitive Neuroscience*, The Free Press/Macmillan, New York, 1993.
- [13] Pollard, C. and Sag, I. *Head-driven Phrase Structure Grammar*, CSLI, Stanford, CA, in press [distributed by University of Chicago Press].
- [14] Pulvermüller, F., “Constituents of a neurological theory of language”, *Concepts in Neuroscience*, 3 (2), 1990, 197–200.
- [15] Quine, W.V.O., “Two dogmas of empiricism”, in W.V.O. Quine, *From a Logical Point of View*, Harvard University Press, Cambridge, MA, 1953.
- [16] Quine, W.V.O., *Word and Object*, MIT Press, Cambridge, MA, 1960.
- [17] Quine, W.V.O., “Epistemology naturalized”, in W.V.O. Quine, *Ontological Relativity*, Columbia University Press, New York, 1969.
- [18] Quine, W.V.O., *Pursuit of Truth*, Harvard University Press, Cambridge, MA, 1990.
- [19] Saussure, F. de, *Cours de linguistique générale*, Payot, Paris, 1962.
- [20] Schnelle, H. “Turing naturalized—Von Neumann’s unfinished project”, in R. Herken (ed.), *The Universal Turing Machine—A Half-Century Survey*, Oxford University Press, Oxford, 1987, 539–559.
- [21] Schnelle, H., *Die Natur der Sprache*, de Gruyter, Berlin, 1991.
- [22] Schnelle, H., “Language and brain”, in W. Nöth (ed.), *Origins of Semiosis*, de Gruyter-Mouton, Berlin, 1994.

- [23] Schnelle, H., Doust, R., “A net-linguistic Earley-Parser”, in R. Reilly, N.E. Sharkey (eds.), *Connectionist Approaches to Languages, Vol. 1*, Lawrence Erlbaum, Hillsdale, NJ, 1992.
- [24] Sejnowski, T.J., Churchland, P.S., “Brain and Cognition”, in M.I. Posner, (ed.), *Foundations of Cognitive Science*, MIT Press, Cambridge, MA, 1989, 301–356.
- [25] Wilkens, R., Schnelle, H., “A connectionist parser for context-free phrase structure grammars”, in G. Dorffner (ed.), *Konnektionismus in Artificial Intelligence und Kognitionsforschung*, Springer, Berlin, 1990, 38–47.

### **SECTION 3**

#### **THE ACQUISITION AND USE OF LARGE CORPORA**

# The Ecology of Language\*

Donald E. Walker

Artificial Intelligence and Information Science Research  
Bellcore

## Abstract

The *ecology of language* is a concept introduced to clarify the relationships that hold between the use of language by people and the contexts in which those uses take place. Developing an effective *ecology* requires access to large amounts of textual and lexical data. A number of projects are currently underway in the United States, in Europe, and in Japan to provide such materials in a form that allows them to be shared as community resources. This paper first briefly reviews a number of these efforts. Then it presents the results of a Workshop on Open Lexical and Textual Resources that was convened to discuss the value of amassing such collections. Consideration is subsequently given to types of corpora and procedures for designing and analyzing them. Three projects are singled out for special mention: (1) the *Text Encoding Initiative*, which is formulating and disseminating guidelines for the preparation and dissemination of machine-readable texts for research and for use by the language industries; (2) the *Data Collection Initiative*, which is acquiring and preparing a large text corpus to be made available for scientific research at cost and without royalties; and (3) the *Consortium for Lexical Research*, which is collecting and disseminating lexical resources and providing a clearinghouse for research results that make use of them. Understanding the *ecology of language* and analyzing the resources that contribute to it will prove increasingly important for electronic document delivery.

## 1 The Ecological Analysis of Language<sup>1</sup>

*Ecology* is a term much used in contemporary society to refer to a wide variety of relationships between organisms and their environments. When I apply the term to human language and refer to the *ecology of language*, I am considering in particular the relationships between the use of language by people and the contexts in which those uses take place. I believe that this concept will prove to have critical importance for dictionary development, for the collection and analysis of text corpora, and for computational linguistics more generally.

\*An earlier version of this paper appeared in *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031, February 1991, Japan Electronic Dictionary Research Institute, Tokyo, Japan, 10-22.

<sup>1</sup>More extended descriptions of some of the topics considered in this paper from the standpoint of "Developing Lexical Resources" can be found in Walker (1989). Walker (1990) considers these issues in relation to text collections more generally.

Most research in linguistics and computational linguistics has approached language either abstractly, by analyzing formal models of linguistic structure, or concretely, by selecting a particular instantiation of some linguistic phenomena and studying their specific properties. Theoretical work, influenced by Chomsky, has tended to concentrate on competence models and language universals and to ignore variability in performance. In contrast, the emerging applications dimension in natural language processing has resulted in narrow systems with limited grammars that function effectively only in limited contexts. There is no satisfactory way to compare such systems.<sup>2</sup>

People talk and write differently; they are exposed both aurally and visually to many different kinds of language material. The same person actually uses different kinds of languages in different situations. Consequently, in trying to aggregate bodies of language material for analysis or to choose the correct lexical and textual resources for comparison, we must be conscious of these differences. In particular, in order to proceed systematically, it is necessary to have vast amounts of material, appropriately diversified to represent the different contexts.

There are many archives and collections of text. Most specialize in material of a particular type. Others, like the Oxford Text Archive (Proud 1989) accept a variety of kinds of data, but, since there is no policy guiding acquisition, the coverage is not systematic. The Brown Corpus (Kucera & Francis 1967) was one of the first attempts to collect "samples" of published material with the intention of drawing some conclusions about American English usage: 500 sets of 2000 words gathered from a range of media all published in a particular year in the early 1960s. The Lancaster-Oslo/Bergen (LOB) Corpus (Johansson 1985) contains similar coverage for British English for that same year. More recently, the Birmingham University Corpus (Sinclair 1987b) was collected to include larger amounts and a broader range of British English. It is of particular interest since a substantial part of its 20 million words were used in the development of the COBUILD Dictionary (Sinclair 1987a). Another collection that has been much studied is the London-Lund Corpus (Svartvik & Quirk 1980) for spoken English. These and other material from a variety of languages have been useful resources for research. However, it has become increasingly clear that they are not representative.

Similar observations can be made about dictionaries. Even the large unabridged dictionaries do not cover technical terms, so separate volumes are necessary for law, medicine, and other specialty fields. Moreover, most dictionaries presume that the user will provide the context for interpretation. Learner's dictionaries are an exception; here an attempt is made to specify where a given word or expression should be used. However, the data necessary to provide proper guidance are not yet available. Strong interactions between textual and lexical research in an ecological perspective are essential.<sup>3</sup>

To establish that a collection is representative, it is necessary to determine what

---

<sup>2</sup>Palmer and Finin (1990) present the results of a "Workshop on Evaluation of Natural Language Processing Systems" that took place in December 1988. They remark on the problems posed by the lack of a common corpus of annotated language. A second workshop is scheduled for June 1991 in conjunction with the 1991 Annual Meeting of the *Association for Computational Linguistics*.

<sup>3</sup>See Walker (1986, 1987) for an analysis of the interaction between *sources*, that is, large document collections, and *resources*, that is, tools like machine-readable dictionaries, encyclopedias, and other specialized guidebooks. There is a reciprocal relation in which the resources provide the basis for more refined access to the information contained in the sources, while the source materials are processed to derive additional resource data.

population is being sampled. A critical requirement is the development of a methodology. There are many perspectives from which the population issue can be approached. One extreme would be to consider the reference population to consist of all the language that has ever been written or spoken. Although that is unrealizable, it does provide an upper bound on what we might consider relevant for inclusion. Another possibility, of particular interest to psycholinguists, would be all the language to which an individual has been exposed during his or her lifetime. The complement is all the language that person has produced. For both, it would be difficult to be comprehensive, but it is possible to conceive of a longitudinal study that would provide samples. A third would be all the language a particular individual produces or is exposed to during a given period of time, say over the course of a “typical” day. From the vantage point of sublanguages or genres, one could consider all the material in some area, for example, classical Greek or Latin, or epic poetry, although the body of literature on computational linguistics itself might also be considered. Each of these alternatives entails a different strategy for collection and different levels of confidence in the adequacy of the sample with respect to its population.

## 2 Developing Resources for Natural Language Research

The development of an effective ecology of language requires access to large amounts of resources of various kinds. We need standards for representing the form and content of texts and lexical data. We need both textual and lexical materials in large quantities, and we need new procedures to analyze them. Fortunately, there are a number of very exciting activities in progress that will provide at least some of these resources. I present a brief selective review here. In later sections I will elaborate on a few of the these activities that I have been involved with more directly. Some of the others are discussed in other papers at this Workshop.

### Text Encoding Initiative

The *Text Encoding Initiative (TEI)* is a major international project established to develop guidelines for the encoding and interchange of machine-readable text and related data. It builds on the Standard Generalized Markup Language (SGML),<sup>4</sup> which is a way of characterizing the content of a document descriptively rather than specifying the procedures needed to convert it to a particular print or display form. In particular, the goal of the TEI is to encourage the development of “tag sets” for representing the features contained in a wide variety of text types. Having standard representations for dictionaries and large text files will make it easier to share data and to compare and evaluate the effectiveness of different algorithms for processing them. Because of its importance, because the Association for Computational Linguistics is one of the sponsoring organizations, and because I am a member of the TEI Steering Committee, I will talk about it in more detail in a subsequent section.

---

<sup>4</sup>ISO Standard 8879 (International Standards Organization 1986). Appropriate references are Bryan (1988), van Herwijnen (1990), and Goldfarb (1990).

## **ACL Data Collection Initiative**

The next four items involve data collection. I begin with the *Data Collection Initiative* sponsored by the Association for Computational Linguistics (ACL/DCI), again because I have been so closely associated with it. Here the aegis of a not-for-profit scientific society is being used to oversee the acquisition and preparation of a large text corpus to be made available for scientific research essentially at the cost of reproduction and without royalties. The data would be converted into a standard SGML form, and progressively tagged and annotated to identify consensually approved linguistic features. In the early stages of the project, we have been accepting material of all kinds without respect for sampling or population source. Subsequently, the ecological issues of sampling and representativeness will be addressed. While most of the current contents reflect American English, there are sets of bilingual and multilingual texts, not all including English as one of the languages. The ACL/DCI will be described in more detail in a subsequent section.

## **British National Corpus Project**

The *British National Corpus* is another collection effort that has just been established to provide a representative sample of British English. It follows in the tradition of the Brown, LOB, and Birmingham corpora referred to above. Oxford University Press, the Longman Group, Oxford University, the University of Lancaster, and The British Library, are cooperating in this effort.<sup>5</sup>

## **Network of European Textual Corpora**

A similar but more comprehensive project is being undertaken by the European Community to coordinate the collection of texts for the major languages within the European Community. The *Network of European Textual Corpora* will use a common methodology for collecting materials; will establish shared standards for coding them, based on the work of the *Text Encoding Initiative*; and will develop software that can be used for all the different language sources.

## **North American Center for Machine-Readable Texts in the Humanities**

A more sharply focused text cataloguing and collection effort has been established by Rutgers and Princeton Universities. They are cooperating in the formation of a *Center for Machine Readable Texts in the Humanities*. Its goals are to: create an on-going inventory of machine-readable texts in the humanities; catalog these items in an online database and make their existence known internationally; acquire, preserve, and service humanities textual data files that would otherwise disappear; make the files available to scholars in the most convenient and least expensive manner; and act as a resource center

---

<sup>5</sup>Della Summers (1991), in her presentation at this Workshop of the work going on at Longman, described the "Longman Lancaster Corpus," a 30-million-word collection of British, American, and other varieties of English that is being used for lexicographic purposes, but which is also being made available to scholars for research.

and referral point for humanities scholars to other projects and centers that disseminate humanities textual data.

## **ESPRIT ACQUILEX**

Shifting now to the lexical arena, I will begin by reviewing some European efforts first, concluding with the *Japanese Electronic Dictionary Research Project*, which is the focus of the present Workshop. Because it was the first effort to coordinate European research in this area, it is appropriate to start by considering the *ACQUILEX* project carried out by the European Community under its ESPRIT program to develop procedures for extracting data from machine-readable dictionaries. Its objective was to modify techniques used for single monolingual dictionaries so they could be used to construct a single integrated multilingual lexical knowledge base from multiple dictionary sources in different languages.

## **EUROTRA-7 and ESPRIT MULTILEX**

*EUROTRA-7* was a feasibility study sponsored by the European Community to determine the reusability of lexical and terminological resources. The positive results from that effort led to the establishment of *MULTILEX*, an ESPRIT project intended to propose a standard for a European multilingual and multifunctional lexicon, to develop a lexical database according to the standard, to create an integrated and robust set of tools to store and retrieve lexical information, and to test the resulting lexicon in prototypical applications.<sup>6</sup>

## **EUREKA GENELEX**

*GENELEX* is a EUREKA project established to construct comprehensive generic dictionaries in French, Italian, and Spanish based on a common data model that is intended to become a standard for lexical representation. It will also develop sets of software to merge existing dictionaries into the *GENELEX* model, to parse texts to update the generic lexicon, to create a lexicographer's workstation to administer the database, and to select and generate sublexicons for particular applications. The final objective is to demonstrate the utility of the generic dictionaries for practical industrial applications.<sup>7</sup>

## **Japanese Electronic Dictionary Research Project**

The *Japanese Electronic Dictionary Research Project*, which this Workshop is celebrating, is the largest and most comprehensive effort yet made to develop a global electronic dictionary. There are several distinctive features that should be highlighted: (1) it is working simultaneously with two languages, Japanese and English, for each of which are being developed 200,000 word general vocabularies and 100,000 word terminological

---

<sup>6</sup>John McNaught (1991) in a paper on "Reusability of Lexical and Terminological Resources: Steps towards Independence," presented at this Workshop, provides a detailed review of both projects.

<sup>7</sup>Bernard Normier and Marc Nossin (1991) in a paper on "GENELEX Project: EUREKA for Linguistic Engineering" presented at this Workshop describe this work.

vocabularies; (2) it is creating a 400,000 item concept dictionary that provides both classifications and descriptions; (3) it is providing 300,000 word co-occurrence dictionaries for each language that identify surface collocations; (4) it is including Japanese-English and English-Japanese bilingual dictionaries, each containing 300,000 words; (5) and it is building on a large text corpus to extract dictionary data.<sup>8</sup>

## Consortium for Lexical Research

A *Consortium for Lexical Research* is just being established in the United States at New Mexico State University under the auspices of the Association for Computational Linguistics. The objective is to provide a repository of lexical data and tools that would be shared by universities, publishers, and research institutions as a “pre-competitive lexical resource” for natural language processing systems. People would withdraw resources to perform research, and, if appropriate, return the results of the research to the Consortium for others to share. I will describe this work in more detail later.<sup>9</sup>

## Penn Treebank Project

The *Treebank Project* at the University of Pennsylvania is only one of a number of efforts to annotate written and spoken texts with information on linguistic structure.<sup>10</sup> I single it out for mention here because it is coordinated with the ACL Data Collection Initiative and thus illustrates the interaction between collecting and analysis. The texts used in the Treebank Project are taken from the ACL/DCI and the resulting tagged materials are returned to the ACL/DCI as additional resources for use in other research activities. The objective of the project is to annotate millions of sentences with part-of-speech assignments, skeletal syntactic parsing, intonational boundaries for spoken materials, and other forms of linguistic data that can be encoded consistently and quickly. Although human experts are validating the codings, tools like Church’s (1988) stochastic parser and Hindle’s (1983, 1989) deterministic parser provide candidate structural analyses.

### 2.0.1 Survey of Language Data in Machine-Readable Form

The availability of written and spoken language data in machine-readable form has become a critical requirement for research and development in computational, theoretical, and applied linguistics, in lexicology and lexicography, in literary and humanistic computing, in information retrieval and terminology, and in the language industries more generally. Recognizing the need for a comprehensive inventory of such materials worldwide, a number of organizations are participating in the distribution of a *Survey of*

---

<sup>8</sup>A number of papers at this Workshop describe the EDR Project: “Electronic Dictionary” by Hiroshi Uchida (1991); “How to Define Concepts for Electronic Dictionaries” by Seiji Miike (1991); “How to Extract Dictionary Data from the EDR Corpus” by Yoshio Nakao (1991); and “How to Organize a Concept Hierarchy” by Eiji Yokota (1991).

<sup>9</sup>A presentation by Louise Guthrie made at this Workshop also described the CLR (Wilks and Guthrie 1991).

<sup>10</sup>Garside, Leech, and Sampson (1987) describe the work done on the annotation of the LOB Corpus.

*Language Data in Machine-Readable Form.* The objective is to provide a detailed catalogue of text collections and corpora, machine-readable dictionaries and other lexical resources, and speech data.<sup>11</sup>

### **ATR Automatic Telephone Interpretation System**

It is appropriate to conclude this brief overview of major projects to collect natural language resources by mentioning the *ATR Automatic Telephone Interpretation System*, which is being designed to provide two-way spoken language interpretation over telephones between persons speaking different languages (Kurematsu 1990). It combines technologies for speech recognition, machine translation, and speech synthesis. These technologies, in turn, require research contributions from groups working on all the activities described above.

## **3 A Workshop on Open Lexical and Textual Resources**

The activities presented in this section constitute major contributions to the creation of research resources that are relevant for the ecology of language. However, since the items listed are only a sample of those actually underway, one of the problems is how to coordinate efforts so duplication can be minimized and the results can be shared. To address these issues, Mark Liberman and I organized a workshop on "Open Lexical and Textual Resources" at the University of Pennsylvania in October 1990 under the auspices of the Association for Computational Linguistics. This workshop was motivated by our interest in evaluating the need for computational dictionaries, in establishing the value of large text files for creating those dictionaries and for a range of other purposes, and in determining how cooperation would further the development of these resources. It was funded by the United States National Science Foundation.

The people invited to the workshop represented a number of scientific communities: message and document understanding and retrieval, speech recognition and synthesis, machine translation and multilingual corpora, lexicography and lexicology, psychology, linguistics, and computational linguistics more generally. There were academic, commercial, and governmental participants. Although most were from the United States, there were also representatives from Europe and Japan.

Most of the conclusions reached by the workshop were clear and unambiguous: (1) there is a definite need for massive text collections in standard formats; (2) large subsets of text must be tagged and annotated with linguistic structural features and other information; (3) efficient software tools are required for text management; (4) larger and more detailed machine-readable dictionaries are essential; (5) lexical databases and knowledge bases need to be established; (6) cooperation and support from the publishing community is critical; (7) support from national and regional governments is desirable both to make their own documentation available for research and for help in dealing with the copyright problem. The one area where there was disagreement was about the possibility of developing a single comprehensive lexicon that would be useful for

---

<sup>11</sup>To get a copy of the survey, contact Mrs. Kick Sprangers, INK International, PO Box 75477, NL-1070 AL Amsterdam, NETHERLANDS; +31-20-164591 phone; +31-20-163851 fax.

a broad range of applications. This question is particularly appropriate to raise at the present Workshop, because it constitutes a challenge to the Electronic Dictionary Research Project: to what degree can people developing and using different natural language processing systems actually share lexical data? I believe that a lot of sharing can take place, but it may be necessary to extract different sets of material from a global dictionary for special applications.

A particularly gratifying result of the workshop was the fact that people representing so many different fields of interest in both the commercial and academic communities expressed willingness to cooperate. Equally important is the recognition by national and regional governments of the importance of textual and lexical resources as part of the scientific infrastructure and their need to have more interactions at national and regional levels.<sup>12</sup>

## 4 Types of Corpora

It is generally acknowledged that there are different types of text collections, but the terminology used to characterize them is not consistent, and, in fact, the distinctions are not quite clear. Since the establishment of a typology is important for ecological studies, I will attempt one here, recognizing that it will not be a definitive statement. Moreover, there are clearly examples that bridge several types.

### Heterogeneous

The simplest kind of collection is one that acquires a broad range of different kinds of materials without any well specified criteria for selection. The ACL/DCI is a clear example. During its initial stages, we have been trying to gather as much material of as many different kinds as possible for use by the computational linguistics research community (although not restricted to that). Our initial distribution will provide a subset, but not one based on any systematic sampling procedures.

Another example is provided by the Oxford Text Archive (Proud 1989), which has been accumulating electronic versions of literary and linguistic texts for the past 15 years.<sup>13</sup> It has approximately 1000 titles that occupy a gigabyte of storage in 20 different languages. Items vary in format and quality, and they are redistributed exactly in the form received.

### Homogeneous

The most obvious contrast with "heterogeneous" is "homogeneous." An example is provided by the TIPSTER project of the United States Government. People have been collecting large quantities of messages that occur in military situations, reflecting equipment failure, damage assessment, and the like. More recently, they have shifted

---

<sup>12</sup>The paper by Toshio Yokoi (1991) on "Collaboration and Cooperation for Development of Electronic Dictionaries" at this Workshop further testifies to the need for and interest in sharing resources on a global level.

<sup>13</sup>For further information, contact the Oxford Text Archive, OUCS, 13 Banbury Road, Oxford OX2 6NN, UK; +44-865-273238 phone; +44-865-273275 fax; email archive@vax.oxford.ac.uk.

their focus to newspaper stories. Of particular interest for this Workshop in Japan is the fact that they are collecting both English and Japanese materials. The primary motivation for doing that is to insure that the software they develop for analyzing message content will be applicable to different languages.

### **Systematic**

Systematic collections are designed to provide representative sets of materials. The Brown Corpus (Kucera and Francis 1967) was one of the earliest collections. As noted above, it provided 2000 word extracts from 500 different kinds of texts published during a single year in the United States. The Lancaster/Oslo/Bergen Corpus (Johansson 1985) provided a comparable body of data for British English for that same period. The British National Corpus and the Network of European Textual Corpora referred to above are also examples of this type, although they will use much larger amounts of data for their elements and can be expected to consider frequency as a basis for determining those amounts.

### **Specialized**

There are a variety of specialized collections. The *Thesaurus Linguae Graecae* (Brunner 1987) contains all the known material in Classical Greek. From the standpoint of language, the basis for selection is homogeneous; however, the contents are varied, since it includes essays, histories, plays, poetry, and inscriptions of all kinds. In the United States, LEXIS and WESTLAW contain massive amounts of legal information. Similar databases are available for medicine, MEDLINE and MEDLARS being the most comprehensive. The North American Center for Machine-/Readable Texts in the Humanities, referred to above, is an example for the humanities.

## **5 Designing and Analyzing Corpora**

The proper design and analysis of corpora is a central problem in the ecology of language, and a challenge for future research. However, there has been a lot of interesting work already done on which we can build. In this paper I can only provide a few pointers to the literature. First it is appropriate to note that sociolinguists have investigated these problems extensively (Hymes 1974). Indeed, the well-established concepts of dialect, genre, style, register, and the like can take on a new and particularly significant perspective in a computational context. Recent work by Biber (1988, 1989, 1991) is of particular interest in this regard. He remarks, quite properly, that "a typology of texts is a research prerequisite to any comparative register analysis, whether of speech and writing, formal and informal texts, restricted and elaborated codes, literary and colloquial styles, 'good' and 'bad' student compositions, early and late historical periods, or whatever, to situate particular texts relative to the range of texts in English" (Biber 1989, p. 4). His typology, based on a multivariate analysis of sets of lexical and syntactical features that co-occur in texts, begins with the identification of five dimensions of variation: involved vs. informational production, narrative vs. nonnarrative concerns, explicit vs. situation-dependent reference, overt vs. covert expression of persuasion, and abstract

vs. nonabstract style. A factor analysis results in the identification of eight classes: intimate interpersonal interaction, informational interaction, scientific exposition, learned exposition, imaginative narrative, general narrative exposition, situated reportage, and involved persuasion. These labels do not do full justice to the kinds of materials each subsumes, but they do demonstrate clearly that the socially defined categories like fiction and nonfiction or newspapers and magazines are not adequate. At the same time, the types Biber identifies are constrained by the particular linguistic features (67) he worked with and, of course by the spoken and written texts (481) he selected from the Lancaster-Oslo/Bergen and London-Lund Corpora. In a recent paper, he has been examining the implications of his analysis for the issue of "Representativeness in Corpus Design" (Biber 1991). Here, he considers much more fully the questions of population and sampling.

Atkins, Clear, and Ostler (1991) provide another perspective on corpus design that addresses a broader range of criteria associated with stages in corpus building, copyright issues, markup, and potential users and uses, in addition to typology, population, and sampling. Much more work along these lines is needed.

Another body of research that is relevant is the study of sublanguages. Kittredge and Lehrberger (1982) and Grishman and Kittredge (1986) provide examples of the work in this area. Within this framework, Walker and Amsler (1986) showed how the subject codes of the *Longman Dictionary of Contemporary English* (Procter 1978), which characterize different senses of a given word, could be used to identify the topics of stories from the *New York Times* newswire service. While those codes are useful for "general" English, it is clear that extensive studies of the occurrences of vocabulary in specialized subject areas will be necessary to develop a broader, deeper, and more refined subject-coded scheme.

## 6 The Text Encoding Initiative

As I noted above, one of the first requirements for working with large textual and lexical files is to have the material in a standard form. This concern was expressed strongly in 1987 by a group of scholars in the humanities, reacting to the heterogeneity of coding schemes used in their field, but the implications have resonated much more broadly. The result was the establishment of the *Text Encoding Initiative (TEI)* to formulate and disseminate guidelines for the preparation and interchange of machine-readable texts for scholarly research and to satisfy a broad range of uses in language engineering and by the language industries more generally. It is appropriate to note that the term "text" is interpreted quite broadly here. The *TEI* is equally concerned with equations, figures, tables, spoken language, diagrams, images, and hypertext linkages.

The *TEI* is sponsored by the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. It is funded by the U.S. National Endowment for the Humanities, the European Community, and the Mellon Foundation. A number of affiliated projects are serving as test beds for evaluating the standards as they are developed. More than a dozen professional societies are pledged to endorse and promote the guidelines when they are completed.

The TEI is guided by a Steering Committee consisting of representatives from the three sponsoring organizations. In addition, there are four major Committees and an increasing number of Working Groups.

The *Committee on Text Documentation*, with a membership drawn largely from the library and archive management communities, is developing procedures for the cataloguing and identification of machine-readable texts. In addition to the standard bibliographical information, it includes information on the particular electronic features that are encoded.

The *Committee on Text Representation* is encoding character sets, page layout, and other features physically represented in the source material. It is providing precise recommendations for those textual features for which a convention already exists in printed or written sources. Within this Committee a number of Working Groups have been or are being established to address specific problem areas: character sets; formulae, tables, and related materials; hypertext and hypermedia structures; text criticism; physical description; analytic bibliography; language corpora; reference works.

The *Committee on Text Analysis and Interpretation* is providing discipline-specific sets of tags appropriate to the analytic procedures favored by that discipline, but in such a way as to permit their extension and generalization to other disciplines using analogous procedures. Much of the work in the first phase of the project concentrated on determining how to represent linguistic structure. Of particular interest from the standpoint of the present Workshop is work on dictionary encoding (Amsler and Tompa 1988). The Working Groups recently established to continue and extend the work of this Committee include linguistics, spoken texts, literary studies, historical studies, dictionaries, natural language processing lexicons, and terminology. Others are likely to be added.

The *Committee on Syntax and Metalanguage* has determined that the syntactic framework of SGML is adequate for foreseeable applications within the scope of the TEI and thus will provide the basic syntax. However, it is suggesting a number of modifications which will be presented to the International Standards Organization. This committee is now surveying major existing schemes and developing a formal metalanguage that can be used to describe both these schemes and one being developed for the Guidelines. The metalanguage can provide formally specifiable mappings between pairs of schemes.

The first two-year development cycle of the TEI was completed in June 1990. The results are embodied in a report, "Guidelines for the Encoding and Interchange of Machine-Readable Texts" (Sperberg-McQueen and Burnard 1990), which has been widely disseminated for review and comment. It considers the general issue of SGML markup, character sets, documentation and bibliographical control, features common to many text types, analytic and interpretive features with particular emphasis on common constructs for linguistic analysis, features for specific text types (office documents, language corpora, dictionaries), and other relevant information. It is essential to have as many people as possible read, test, and evaluate this document.<sup>14</sup>

---

<sup>14</sup>To obtain copies of the report and for further information, contact the editors: C.M. Sperberg-McQueen (Computer Center (M/C 135), University of Illinois at Chicago, Box 6998, Chicago, IL 60680, USA; +1-312-9962477; u35395@uicvm.cc.uic.edu, u35395@uicvm.bitnet) or Lou Burnard (Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, ENGLAND; +44-865-273238, 273200; lou@vax.ox.ac.uk). Public information is provided through TEI-L on the bitnet LISTSERV node at UICVM.

During the second two-year cycle, a range of affiliated projects will apply the current guidelines, while the Committees and Working Groups will refine and extend them. A final report is scheduled to appear in July 1992.

## 7 The ACL Data Collection Initiative

The Association for Computational Linguistics established its Data Collection Initiative (ACL/DCI) in response to a recent upsurge of interest in computational studies of large bodies of text. The aim of such studies varies widely, from lexicography and studies of language change to automatic indexing methods and statistical models for improving the performance of optical character readers. In general, corpus-based studies are essential for the development of adequate models of linguistic structure and for insights into the nature of language use. However, researchers have been severely hampered by the lack of appropriate materials, and specifically by the lack of a large enough body of text on which published results can be replicated or extended by others.

As noted in the brief description of the ACL/DCI in the section reviewing current activities, the objective is to make a large text corpus available for scientific research essentially at cost and without royalties. The initial goal was to acquire at least 100 million words, but much more than that amount has already been received and even more has been pledged. Either the materials are out of copyright, or permissions from the copyright owners have been granted that allow us to include them. Each applicant for data from the ACL/DCI is required to sign an agreement not to redistribute the data or use it for other than research purposes. Special restrictions on some materials may be necessary, but that will only be done if the restrictions do not compromise the central objective of providing general long-term access for research.

Processing the acquired data for inclusion in the ACL/DCI requires removing special tape record structures, mapping other character codes into ASCII, interpreting or eliminating special typographical or formatting codes, and adding high-level markup in SGML to express structural features of the document, like font definition, point size, or page layout (as characterized by the TEI Committee on Text Representation). Subsequently we intend to add more analytic information like part-of-speech assignments, skeletal syntactic parsings, intonational boundaries, and other forms of linguistic structure that can be encoded rapidly and reliably (of the kind being specified by the DTEI Working Group on Linguistics of the Committee on Text Analysis and Interpretation, and exemplified by the work of the Penn Treebank Project).

The initial distribution, which will be on CD-ROM, will include material from the Wall Street Journal, the Canadian Hansard parliamentary records, the Library of America publications of classic American literature and history, transcripts of the Challenger commission hearings, Department of Energy scientific abstracts, fact sheets from the U.S. Department of Agriculture Extension Service, and other textual data.<sup>15</sup>

---

Contact Sperberg-McQueen to have your name added to the list.

<sup>15</sup>For further information, contact Mark Liberman (DCI), Linguistics Department, 619 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104, USA; +1-215-8986046.

## **8 The Consortium for Lexical Research**

The Consortium for Lexical Research (CLR) is the lexical counterpart of the ACL/DCI. As noted above in the overview of activities, the objective is to provide a shared repository of lexical data and tools. Participants would contribute their own materials and would have the right to withdraw resources in order to carry out specific research projects. The results of the research, in turn, would be contributed to the repository in the form of data, tools, and theoretical insights. Originally proposed by Roy Byrd of the IBM T.J. Watson Research Center, a committee of the Association for Computational Linguistics was formed to oversee its development. It has now been established at New Mexico State University with funding by the Defense Advanced Research Projects Agency of the United States Government.

Lexicon construction is a costly and time consuming task, as the experience of the Japan Electronic Dictionary Research Project demonstrates. Moreover, since each natural language processing system developer needs lexical information, there can be a substantial duplication of effort if each organization proceeds independently. The CLR is intended to encourage cooperation in the development of the data everyone needs and that does not constitute a competitive advantage.

The following types of lexical data are among those that are of interest: published dictionaries as typesetting tapes and lexical databases; lists of words, phrases, collocations, and idioms—incorporating phonological, morphological, syntactic, semantic, and prosodic information, if available; proper nouns and specialized terminologies and glossaries; synonomous and other relations among word and phrases; statistical data on frequency of occurrence.

The software tools would include: database management and access procedures for lexical databases; concordance and related programs for analyzing text corpora; morphological analyzers; parsers for determining part-of-speech and other syntactic information; and tools for converting and processing SGML data, particularly in the format being developed by the Dictionary Encoding Initiative.

Protecting the intellectual property rights of the participants is a critical concern, because some of the needed materials, like machine-readable dictionaries, are commercial products of the publishers and contain information they would not like to share with their competitors. License agreements may be required, some of which might deal with control of derivative products as well. Special agreements could of course be negotiated between contributors and prospective users that would be outside the CLR, but facilitated by its existence.

In the summary section of the original proposal, Byrd argues that creating a Consortium for Lexical Research would benefit the Natural Language Processing Community “by providing a focal point for the creation of an interacting community of lexical researchers, fostering cooperation on the creation of better lexical resources faster, establishing a center of competence in lexical research, encouraging the creation of reusable lexical tools, techniques, and data, enabling research that has more impact than is possible with simple publication of theoretical results, and removing the necessity to negotiate multiple bilateral joint-study agreements covering this kind of research.” That would indeed be a significant development.

## 9 Conclusions

In characterizing the *ecology of language* at the beginning of this paper, I described it as “the relationships between the use of language by people and the contexts in which those uses take place.” In my remarks I have concentrated on developing lexical and textual resources for natural language research that would support ecological studies. However, I have not dealt at all with the “people” or the “contexts” or the “uses” directly. I believe this delimitation is appropriate at the present stage of our understanding. The “relationships” are the primary data; as we comprehend them, we can begin to analyze their constituent structure.

One of the challenges I face in trying to provide evidence for the value of the concept of the *ecology of language* is a demonstration of its utility. Those of us in the research community believe that machine-readable dictionaries and massive text collections are valuable resources, but we have to demonstrate that value to the outside world. We need a marketplace, things to sell in it, and buyers. Fortunately, an applications component is emerging in computational linguistics. As it does, it can serve as a counterpoint to our theories, so that we can begin to benefit from the interplay of science and engineering that has been so productive in other fields of knowledge.

As a final note, I would like to point out some other *ecologies* that are relevant for our concerns. I recently discovered a book on *The Ecology of Computation* (Huberman 1988) that raises similar questions for computer science. It takes its point of departure from distributed computational systems which the editor believes are acquiring features characteristic of social and biological organizations. Of particular interest is the final chapter on “The Next Knowledge Medium” (Stefik 1988) in which the issues of standardization, collection, and reusability figure in the development of knowledge bases and a knowledge medium. It is possible to generalize his thesis and suggest the need for an *ecology of knowledge*. An appropriate counterpart would also be an *ecology of information*. That is, we need to understand the classes of information that are embodied in our text collections and in our lexical databases and the knowledge that can be said to underlie them. Language, information, and knowledge are all combined in human efforts after communication, so it may be reasonable to talk about an *ecology of communication* as well.

## References

- [1] Amsler, Robert A.; and Tompa, Frank William. 1988. “An SGML-based Standard for English Monolingual Dictionaries.” In *Information in Text: Proceedings of the 4th Annual Conference of the UW Centre for the New Oxford English Dictionary*. University of Waterloo Centre for the New Oxford English Dictionary, Waterloo, Ontario. Pages 61–79.
- [2] Atkins, Sue; Clear, Jeremy; and Ostler, Nicholas. 1991. “Corpus Design Criteria.” Prepared for the Workshop on European Textual Corpora. Manuscript.
- [3] Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, England.

- [4] Biber, Douglas. 1989. "A Typology of English Texts." *Linguistics* 27, 3–43.
- [5] Biber, Douglas. 1991. "Representativeness in Corpus Design." Chapter in present volume.
- [6] Brunner, Theodore F. 1987. "Data Banks for the Humanities: Learning from Thesaurus Linguae Graecae." *Scholarly Communication* 7, 1, 6–9.
- [7] Bryan, Martin. 1988. *SGML: An Author's Guide to the Standard Generalized Markup Language*. Addison-Wesley, Wokingham, England and Reading, Massachusetts.
- [8] Church, Kenneth Ward. 1988. "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text." *Proceedings of the Second Conference on Applied Natural Language Processing*. Austin, Texas, 9–12 February 1988. Association for Computational Linguistics, Morristown, New Jersey, 136–143.
- [9] Garside, Roger; Leech, Geoffrey; and Sampson, Geoffrey (Editors). 1987. *The Computational Analysis of English: A Corpus-Based Approach*. Longman, London and New York.
- [10] Goldfarb, Charles. 1990. *The SGML Handbook*. Oxford University Press, Oxford.
- [11] Grishman, Ralph; and Kittredge, Richard (Editors). 1986. *Analyzing Language in Restricted Domains*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- [12] Hindle, Donald. 1983. "Deterministic Parsing of Syntactic Non-fluencies." *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. Cambridge, Massachusetts, 15–17 June 1983. Association for Computational Linguistics, Morristown, New Jersey, 123–128.
- [13] Hindle, Donald. 1989. "Acquiring Disambiguation Rules from Text." *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, British Columbia, 26–29 June 1989. Association for Computational Linguistics, Morristown, New Jersey, 118–125.
- [14] Huberman, B.A. (Editor). 1988. *The Ecology of Computation*. North-Holland, Amsterdam.
- [15] Hymes, Dell. 1974. *Foundations in Sociolinguistics*. University of Pennsylvania Press, Philadelphia.
- [16] International Standards Organization. 1986. *International Standard ISO 8879: Information Processing—Text and Office Systems—Standard Generalized Markup Language (SGML)*. American National Standards Institute, New York.
- [17] Johansson, Stig. 1985. "Word Frequency and Text Type: Some Observations Based on the LOB Corpus of British English Texts." *Computers and the Humanities* 19:1, 23–26.

- [18] Kittredge, Richard; and Lehrberger, John (Editors). 1982. *Sublanguage: Studies of Language in Restricted Semantic Domains*. de Gruyter, Berlin.
- [19] Kučera, Henry; and Francis, W. Nelson. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, Rhode Island.
- [20] Kurematsu, Akira. 1990. "A Perspective of Telephone Interpretation Research." *Proceedings of Pacific Rim International Conference on Artificial Intelligence '90*, Nagoya, Japan, 14–16 November 1990. Japanese Society for Artificial Intelligence, Tokyo, Japan, 11–16.
- [21] McNaught, John. 1991. "Reusability of Lexical and Terminological Resources: Steps towards Independence." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 97–106.
- [22] Miike, Seiji. 1991. "How to Define Concepts for Electronic Dictionaries." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 43–49.
- [23] Nakao, Yoshio. 1991. "How to Extract Dictionary Data from the EDR Corpus." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 58–62.
- [24] Normier, Bernard; and Nossin, Marc. 1991. "GENELEX Project: EUREKA for Linguistic Engineering." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 63–69.
- [25] Palmer, Martha; and Finin, Tim. 1990. "Workshop on the Evaluation of Natural Language Processing Systems." *Computational Linguistics* 16:3, 175–181.
- [26] Procter, Paul (Editor). 1978. *Longman Dictionary of Contemporary English*. Longman Group Limited, Harlow and London.
- [27] Proud, Judith K. 1989. "The Oxford Text Archive." *British Library R&D Report No. 5985*. British Library, London.
- [28] Sinclair, John M. (Editor). 1987a. *Collins COBUILD English Language Dictionary*. Collins, Glasgow.
- [29] Sinclair, John M. (Editor). 1987b. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins, Glasgow.
- [30] Sperberg-McQueen, C. M.; and Burnard, Lou (Editors). 1990. "Guidelines for the Encoding and Interchange of Machine-Readable Texts." Draft: Version 1.0, 15 July 1990.
- [31] Stefkik, Mark J. 1988. "The Next Knowledge Medium." *The Ecology of Computation*, edited by B.A. Huberman. North-Holland, Amsterdam, 315–342.

- [32] Summers, Della. 1991. "Longman Computerization Initiatives, Corpus Building, Semantic Analysis, and Prolog Version of LDOCE by Cheng-ming Guo." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 141–152.
- [33] Svartvik, Jan; and Quirk, Randolph (Editors). 1980. *A Corpus of English Conversation*. Gleerup, Lund.
- [34] Uchida, Hiroshi. 1991. "Electronic Dictionary." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 23–42.
- [35] van Herwijnen, Eric. 1990. *Practical SGML*. Kluwer, Dordrecht.
- [36] Walker, Donald E. 1986. "Knowledge Resource Tools for Accessing Large Text Files." In *Information in Data: Proceedings of the First Conference of the University of Waterloo Center for the New Oxford English Dictionary*. University of Waterloo Center for the New Oxford English Dictionary Waterloo, Ontario, 11–24.
- [37] Walker, Donald E. 1987. "Knowledge Resource Tools for Accessing Large Text Files." In *Machine Translation: Theoretical and Methodological Issues*, edited by Sergei Nirenberg. Cambridge University Press, Cambridge, England, 247–261.
- [38] Walker, Donald E. 1989. "Developing Lexical Resources." In *Dictionaries in the Electronic Age: Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary*. University of Waterloo Centre for the New Oxford English Dictionary, Waterloo, Ontario, 1–22.
- [39] Walker, Donald E. 1990. "Collecting Texts, Tagging Texts, and Putting Texts in Context." In *Text-Based Intelligent Systems: Current Research in Text Analysis, Information Extraction, and Retrieval*, edited by Paul S. Jacobs. General Electric Research & Development Center Technical Information Series, 90CRD198, Schenectady, New York, September 1990, 30–34.
- [40] Walker, Donald E.; and Amsler, Robert A. 1986. "The Use of Machine-Readable Dictionaries in Sublanguage Analysis." In *Analyzing Language in Restricted Domains*, edited by Ralph Grishman and Richard Kittredge. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 69–83.
- [41] Yokoi, Toshio. 1991. "Collaboration and Cooperation for Development of Electronic Dictionaries." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 204–207.
- [42] Yokota, Eiji. 1991. "How to Organize a Concept Hierarchy." *Proceedings of the International Workshop on Electronic Dictionaries*, EDR TR-031. Japan Electronic Dictionary Research Institute, Tokyo, Japan, 50–57.

# Representativeness in Corpus Design\*

Douglas Biber  
Department of English  
Northern Arizona University  
*e-mail: biber@nauvax.ucc.nau.edu*

## **Abstract**

The present paper addresses a number of issues related to achieving 'representativeness' in linguistic corpus design, including: discussion of what it means to 'represent' a language, definition of the target population, stratified versus proportional sampling of a language, sampling within texts, and issues relating to the required sample size (number of texts) of a corpus. The paper distinguishes among various ways that linguistic features can be distributed within and across texts; it analyzes the distributions of several particular features, and it discusses the implications of these distributions for corpus design.

The paper argues that theoretical research should be prior in corpus design, to identify the situational parameters that distinguish among texts in a speech community, and to identify the types of linguistic features that will be analyzed in the corpus. These theoretical considerations should be complemented by empirical investigations of linguistic variation in a pilot corpus of texts, as a basis for specific sampling decisions. The actual construction of a corpus would then proceed in cycles: the original design based on theoretical and pilot-study analyses, followed by collection of texts, followed by further empirical investigations of linguistic variation and revision of the design.

## **1 General Considerations**

Some of the first considerations in constructing a corpus concern the overall design: for example, the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples. Each of these involves a sampling decision, either conscious or not.

The use of computer-based corpora provides a solid empirical foundation for general purpose language tools and descriptions, and enables analyses of a scope not otherwise possible. However, a corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning a language as a whole; for example, corpus-based dictionaries, grammars, and general part-of-speech taggers are applications requiring a representative basis (cf. Biber 1993b).

---

\*I would like to thank Edward Finegan for his many helpful comments on an earlier draft of this paper. A modified version of this paper was distributed for the Pisa Workshop on Textual Corpora, held at the University of Pisa (January 1992), and discussions with several of the workshop participants were also helpful in revising the paper.

Typically researchers focus on sample size as the most important consideration in achieving representativeness: how many texts must be included in the corpus, and how many words per text sample. Books on sampling theory, however, emphasize that sample size is not the most important consideration in selecting a representative sample; rather, a thorough definition of the target population and decisions concerning the method of sampling are prior considerations. Representativeness refers to the extent to which a sample includes the full range of variability in a population. In corpus design, variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: 1) the range of text types in a language, and 2) the range of linguistic distributions in a language.

Any selection of texts is a sample. Whether or not a sample is ‘representative’, however, depends first of all on the extent to which it is selected from the range of text types in the target population; an assessment of this representativeness thus depends on a prior full definition of the ‘population’ that the sample is intended to represent, and the techniques used to select the sample from that population. Definition of the target population has at least two aspects: 1) the boundaries of the population — what texts are included and excluded from the population; and 2) hierarchical organization within the population — what text categories are included in the population, and what are their definitions. In designing text corpora, these concerns are often not given sufficient attention, and samples are collected without a prior definition of the target population. As a result, there is no possible way to evaluate the adequacy or representativeness of such a corpus (because there is no well-defined conception of what the sample is intended to represent).

In addition, the representativeness of a corpus depends on the extent to which it includes the range of linguistic distributions in the population. That is, different linguistic features are differently distributed (within texts, across texts, across text types), and a representative corpus must enable analysis of these various distributions. This condition of linguistic representativeness depends on the first condition — that is, if a corpus does not represent the range of text types in a population, it will not represent the range of linguistic distributions. In addition, linguistic representativeness depends on issues such as the number of words per text sample, the number of samples per ‘text’, and the number of texts per text type. These issues will be addressed in Sections 3 and 4.

However, the issue of population definition is the first concern in corpus design. To illustrate, consider the population definitions underlying the Brown corpus (Francis and Kučera 1964/1979) and the LOB corpus (Johansson, Leech, and Goodluck 1978). These target populations were defined both with respect to their boundaries (all published English texts printed in 1961, in the U.S. and Great Britain respectively), and their hierarchical organizations (15 major text categories and numerous sub-genre distinctions within these categories). In constructing these corpora, the compilers also had good ‘sampling frames’, enabling probabilistic, random sampling of the population. A sampling frame is an operational definition of the population, an itemized listing of population members from which a representative sample can be chosen. The LOB corpus manual (Johansson, Leech, and Goodluck 1978) is fairly explicit about the sampling frame used: for books, the target population was operationalized as all 1961 publications listed in The British National Bibliography Cumulated Subject Index, 1960-1964

(which is based on the subject divisions of the Dewey Decimal Classification system), and for periodicals and newspapers, the target population was operationalized as all 1961 publications listed in Willing's Press Guide, 1961. In the case of the Brown Corpus, the sampling frame was the collection of books and periodicals in the Brown University Library and the Providence Athenaeum; this sampling frame is less representative of the total texts in print in 1961 than the frames used for construction of the LOB Corpus, but it provided well-defined boundaries and an itemized listing of members. In choosing and evaluating a sampling frame, considerations of efficiency and cost effectiveness must be balanced against higher degrees of representativeness.

Given an adequate sampling frame, it is possible to select a probabilistic sample. There are several kinds of probabilistic samples, but they all rely on random selection. In a simple random sampling, all texts in the population have an equal chance of being selected. For example, if all entries in the British National Bibliography were numbered sequentially, then a table of random numbers could be used to select a random sample of books. Another method of probabilistic sampling, which was apparently used in the construction of the Brown and LOB corpora, is 'stratified sampling'. In this method, sub-groups are identified within the target population (in this case, the genres), and then each of those 'strata' are sampled using random techniques. This approach has the advantage of guaranteeing that all strata are adequately represented while at the same time selecting a non-biased sample within each stratum (i.e., in the case of the Brown and LOB corpora, there was 100% representation at the level of genre categories and an unbiased selection of texts within each genre).

Note that, for two reasons, a careful definition and analysis of the non-linguistic characteristics of the target population is a crucial prerequisite to sampling decisions. First, it is not possible to identify an adequate sampling frame or to evaluate the extent to which a particular sample represents a population until the population itself has been carefully defined. A good illustration is a corpus intended to represent the spoken texts in a language. As there are no catalogs or bibliographies of spoken texts, and since we are all constantly expanding the universe of spoken texts in our everyday conversations, identifying an adequate sampling frame in this case is difficult; but without a prior definition of the boundaries and parameters of speech within a language, evaluation of a given sample is not possible.

The second motivation for a prior definition of the population is that stratified samples are almost always more representative than non-stratified samples (and they are never less representative). This is because identified strata can be fully represented (100% sampling) in the proportions desired, rather than having to depend on random selection techniques. In statistical terms, the between-group variance is typically larger than within-group variance, and thus a sample that forces representation across identifiable groups will be more representative overall.<sup>1</sup> For the Brown and LOB corpora, a prior identification of the genre categories (e.g., press reportage, academic prose, mystery fiction) and sub-genre categories (e.g., medicine, mathematics, and humanities within the genre of academic prose) guaranteed 100% representation at those two levels. That is, the corpus builders attempted to compile an exhaustive listing of the major text

---

<sup>1</sup>Further, in the case of language corpora, proportional representation of texts is usually not desirable (see Section 3); rather, representation of the range of text types is required as a basis for linguistic analyses, making a stratified sample even more essential.

categories of published English prose, and all of these categories were included in the corpus design. Therefore, random sampling techniques were required only to obtain a representative selection of texts from within each sub-genre. The alternative, a random selection from the universe of all published texts, would depend on a large sample and the probabilities associated with random selection to assure representation of the range of variation at all levels (across genres, subgenres, and texts within subgenres) – a more difficult task.

In the present paper, I will first consider issues relating to population definitions for language corpora and attempt to develop a framework for stratified analysis of the corpus population (Section 2). In Section 3, then, I will return to particular sampling issues, including proportional versus non-proportional sampling, sampling within texts (how many words per text and stratified sampling within texts), and issues relating to sample size. In Section 4, I will describe differences in the distributions of linguistic features, presenting the distributions of several particular features, and I will discuss the implications of these distributions for corpus design. Finally, in Section 5, I offer a brief overview of corpus design in practice.

## **2 Strata in a Text Corpus: An Operational Proposal Concerning the Salient Parameters of Register and Dialect Variation**

As noted in the last section, definition of the corpus population requires specification of the boundaries and specification of the strata. If we adopt the ambitious goal of representing a complete language, the population boundaries can be specified as all of the texts in the language. Specifying the relevant strata and identifying sampling frames are obviously more difficult tasks, requiring a theoretically motivated and complete specification of the kinds of texts. In the present section I offer a preliminary proposal for identifying the strata for such a corpus and operationalizing them as sampling frames. The proposal is restricted to western societies (with examples from the United States), and is intended primarily as an illustration rather than a final solution, showing how a corpus of this kind could be designed.

I use the terms genre or register to refer to situationally defined text categories (such as fiction, sports broadcasts, psychology articles), and text type to refer to linguistically defined text categories. Both of these text classification systems are valid, but they have different bases. Although registers/genres are not defined on linguistic grounds, there are statistically important linguistic differences among these categories (Biber 1986, 1988), and linguistic features counts are relatively stable across texts within a register (Biber 1990). In contrast, text types are identified on the basis of shared linguistic co-occurrence patterns, so that the texts within each type are maximally similar in their linguistic characteristics, while the different types are maximally distinct from one another (Biber 1989).

In defining the population for a corpus, register/genre distinctions take precedence over text type distinctions. This is because registers are based on criteria external to the corpus, while text types are based on internal criteria. That is, registers are based on the different situations, purposes, and functions of text in a speech community, and these can be identified prior to the construction of a corpus. In contrast, identification

of the salient text type distinctions in a language requires a representative corpus of texts for analysis; there is no *a priori* way to identify linguistically defined types. As I show in Section 4, though, the results of previous studies, as well as on-going research during the construction of a corpus, can be used to assure that the selection of texts is linguistically as well as situationally representative. For the most part, corpus linguistics has concentrated on register differences.<sup>2</sup> In planning the design of a corpus, however, decisions must be made whether to include a representative range of dialects or to restrict the corpus to a single dialect (e.g., a 'standard' variety). Dialect parameters specify the demographic characteristics of the speakers and writers, including geographic region, age, sex, social class, ethnic group, education, and occupation.<sup>3</sup>

Different overall corpus designs represent different populations and meet different research purposes. Three of the possible overall designs are organized around text production, text reception, and texts as products. The first two of these are demographically organized at the top level. That is, individuals are selected from a larger demographic population, and then these individuals are tracked to record their language use. A production design would include the texts (spoken and written) actually produced by the individuals in the sample; a reception design would include the texts listened to or read. These two approaches would address the question of what people actually do with language on a regular basis. The demographic selection could be stratified along the lines of occupation, sex, age, etc.

A demographically oriented corpus would not represent the range of text types in a language, since many kinds of language are rarely used, even though they are important on other grounds. For example, few individuals will ever write a law or treaty, an insurance contract, or a book of any kind, and some of these kinds of texts are also rarely read. It would thus be difficult to stratify a demographic corpus in such a way that it would insure representativeness of the range of text categories. Many of these categories are very important, however, in defining a culture. A corpus organized around texts as products would be designed to represent the range of registers and text types rather than the typical patterns of use of various demographic groups.

Work on the parameters of register variation has been carried out by anthropological linguists such as Hymes and Duranti, and by functional linguists such as Halliday (see Hymes 1974, Brown and Fraser 1979, Duranti 1985, Halliday and Hasan 1989). In Biber (1993a), I attempt to develop a relatively complete framework, arguing that 'register' should be specified as a continuous (rather than discrete) notion, and distinguishing among the range of situational differences that have been considered in register studies. This framework is overspecified for corpus design work – values on some parameters are entailed by values on other parameters, and some parameters are specific to restricted kinds of texts. Attempting to sample at this level of specificity would thus be extremely difficult. For this reason I propose in Table 1 a reduced set of sampling strata, balancing operational feasibility with the desire to define the target population as completely as

---

<sup>2</sup> Actually, very little work has been carried out on dialect variation from a text-based perspective. Rather, dialect studies have tended to concentrate on phonological variation, downplaying the importance of grammatical and discourse features.

<sup>3</sup> Other demographic factors characterize individual speakers and writers rather than groups of users; these include relatively stable characteristics, such as personality, interests, and beliefs, and temporary characteristics, such as mood and emotional state. These factors are probably not important for corpus design, unless an intended use of the corpus is investigation of personal differences.

---

**Table 1.** Situational parameters listed as hierarchical sampling strata.

---

1. Primary channel: written / spoken / scripted speech
  2. Format: published / not published  
(+ various formats within 'published')
  3. Setting: institutional / other public / private-personal
  4. Addressee:
    - a. plurality: unenumerated / plural / individual / self
    - b. presence (place and time): present / absent
    - c. interactiveness: none / little / extensive
    - d. shared knowledge: general / specialized / personal
  5. Addressor:
    - a. demographic variation: sex, age, occupation, etc.
    - b. acknowledgement: acknowledged individual / institution
  6. Factuality:  
factual-informational / intermediate or indeterminate / imaginative
  7. Purposes: persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship, ...
  8. Topics: ...
-

possible.

The first of these parameters divides the corpus into three major components: writing, speech, and scripted speech. Each of these three requires different sampling considerations, and thus not all subsequent situational parameters are relevant for each component.

Within writing, the first important distinction is publication.<sup>4</sup> This is because the population of published texts can be operationally bounded, and various catalogs and indexes provide itemized listings of members. For example, the following criteria might be used for the operational definition of 'published' texts: 1) they are printed in multiple copies for distribution, and 2) they are copyright registered or recorded by a major indexing service. In the United States, a record of all copyright registered books and periodicals is available at the Library of Congress. Other 'published' texts that are not copyright registered include government reports and documents, legal reports and documents, certain magazines and newspapers, and some dissertations; in the United States, these are indexed in sources such as the Monthly Catalog of U.S. Government Publications, Index to U.S. Government Periodicals, a whole system of legal reports (e.g., the Pacific Reporter, the Supreme Court Reports), periodical indexes (e.g., Readers' Guide to Periodical Literature, Newsbank), and dissertation abstracts (indexed by University Microfilms International).

A third stratum for written published texts could thus be these 'formats' represented by the various cataloging and indexing systems. Together these indexes provide an itemized listing of published writing, and they could therefore be used as an adequate sampling frame. With a large enough sample (see following section), such a sampling frame would help achieve 'representativeness' of the various kinds of published writing. However, we know on theoretical grounds that there are several important substrata within published writing (e.g., purposes and different subject areas), and it is thus better to specify additionally these in the corpus design. This approach is more conservative, in that it insures representativeness in the desired proportions for each of these text categories, and at the same time it enables smaller sample sizes (since random techniques require larger samples than stratified techniques). Setting and format are parallel second-level strata: format is important for the sampling of published writing; setting can be used in a similar way to provide sampling frames for unpublished writing, speech, and scripted speech. Three types of setting are distinguished here: institutional, other public, and private-personal. These settings are less adequate as sampling frames than publication catalogs – they do not provide well-defined boundaries for unpublished writing or speech, and they do not provide an exhaustive listing of texts within these categories. The problem is that no direct sampling frame exists for unpublished writing or speech. Setting, however, can be used indirectly to sample these target populations, by using three separate subcategories: institutions (offices, factories, businesses, schools, churches, hospitals, etc.), private settings (homes), and other public settings (shopping areas, recreation centers, etc.). (For scripted speech, the category of other public settings would include speech on various public media, such as news broadcasts, scripted speeches, and scripted dialogue on television sitcoms and dramas.) Operational sampling frames for each of these settings might be defined from various government and official records (e.g., census records, tax returns, or other registrations). The goal of such sampling

---

<sup>4</sup>This parameter would not be important for many non-western societies, or for certain kinds of corpora representing different historical periods; quite different sampling strategies would be required in these cases.

frames would be to provide an itemized listing of the members within each setting type, so that a random sample of institutions, homes, and other public places could be selected. (These three settings could be further stratified with respect to the various types of institution, types of home, etc.) To this point, then, I have proposed the sampling frames in Table 2 :

Table 2. Outline of sampling frames.

writing:	published:	books / periodicals / etc. (based on available indexes)
	unpublished:	institutional / public / private
speech:	institutional / public / private	
scripted speech:	institutional / public media / other	

Before proceeding, it is necessary to distinguish between two types of sampling strata. The first, as above, actually defines a sampling frame, specifying the boundaries of the operationalized population and providing an itemized listing of members. The second, as in the remaining parameters of Table 1, identifies categories that must be represented in a corpus but do not provide well-defined sampling frames. For example, Addressee plurality (#4a: unenumerated / plural / individual / self) provides no listing of the texts with these four types of addressee; rather, it simply specifies that texts should be collected until these four categories are adequately represented.

Further, the remaining parameters in Table 1 are not equally relevant for all the major sampling frames listed in Table 2. Consider, for example, the parameters listed under 'Addressee'. Published writing always has unenumerated addressees,<sup>5</sup> is always written for non-present addressees, and is almost always non-interactive (except for published exchanges of opinion). It can require either general or specialized background knowledge (e.g., popular magazines versus academic journals) but rarely requires personal background knowledge (although this is needed for a full understanding of memoirs, published letters, diaries, and even some novels and short stories). Unpublished writing, on the other hand, can fall into all of these addressee categories. The addressees can be unenumerated (e.g., advertisements, merchandising catalogs, government forms or announcements), plural (office circular or memo, business or technical report), individual (memo to an individual, professional or personal letter, e-mail message), or self (diary, notes, shopping list). The addressee of unpublished texts is usually absent, except in writing to oneself. Unpublished writing can be interactive (e.g., letters) or not. Finally, unpublished writing can require only general background knowledge (e.g., some advertisements), specialized knowledge (e.g., technical reports), or personal knowledge (e.g., letters and diaries).

Speech is typically directed to a plural or individual addressee, who is present. Speech addressed to self is often considered strange. Speech can be directed to unenumerated, absent addressees through mass media (e.g., a televised interview). Individual

<sup>5</sup>Published collections of letters and published diaries are special cases – these originally have individual addressees, but they are usually written with the hope of eventual publication and thus with an unenumerated audience in mind.

and small-group addressees can also be absent, as in the case of telephone conversations and ‘conference calls’. (Individual addressees can even be non-interactive in the case of talking to an answering machine.) Private settings favor interactive addressees (either individual or small group conversations) while both interactive and non-interactive addressees can be found in institutional settings (for example, consider the various kinds of lectures, sermons, and business presentations). General knowledge can be required in all kinds of conversation; specialized background knowledge is mostly required of addressees in institutional settings; personal knowledge is most needed in private settings.

Scripted speech is typically directed towards plural addressees (small groups in institutional settings and unenumerated audiences for mass media). Dialogue in plays and televised dramatic shows is an example of scripted speech that is directed to an individual but heard by an unenumerated audience. Addressees are typically present for scripted speech in institutional settings but are not present (physically or temporally) for scripted speech projected over mass media. Except for the lecturer who allows questions during a written speech, scripted speech is generally not interactive. Finally, scripted speech can require either general or specialized background knowledge on the part of the addressee, but it rarely requires personal background knowledge.

Addressors can vary along a number of demographic parameters (the dialect characteristics mentioned above), and decisions must be made concerning the representation of these parameters in the corpus. (Collection of texts from some addressor categories will be difficult for some sampling frames; for example, there are relatively few published written texts by working class writers.) The second parameter here, whether the addressor is acknowledged or not, is relevant only for written texts: some written texts do not have an acknowledged personal author (e.g., advertisements, catalogs, laws and treaties, government forms, business contracts), while the more typical kinds of writing have a specific author(s).

Factuality is similar to assessments of background knowledge in that it is sometimes difficult to measure reliably, but this is an important parameter distinguishing among texts within both writing and speech. At one pole are scientific reports and lectures, which purport to be factual, and at the other are the various kinds of imaginative stories. In between these poles are a continuum of texts with different bases in fact, ranging over speculation, opinion, historical fiction, gossip, etc.

The parameter of purpose requires further research, both theoretical (as the basis for corpus design) and empirical (using the resources of large corpora). I include in Table 1 several of the purposes that should be represented in a corpus, but this is not intended as an exhaustive listing.

Similarly the parameter of topic requires further theoretical and empirical research. Library classification systems are well developed and provide adequate topic strata for published written texts. These same classifications might also serve as strata for unpublished writing, but they would need to be tested empirically. For spoken texts, especially in private settings, further research on the range of typical topics is required.

The spirit of the proposal outlined in this section is to show how basic situational parameters can be used as sampling strata to provide an important first step towards achieving representativeness. The particular parameter values used, however, must be refined, and the framework proposed here is clearly not the final word on corpus sampling strata.

### 3 Other Sampling Issues

#### 3.1 Proportional Sampling

In most stratified sample designs, the selection of observations across strata must be proportional in order to be considered representative (Henry 1990, Williams 1978). That is, the number of observations in each stratum should be proportional to their numbers in the larger population. For example, a survey of citizens in North Carolina (reported in Henry 1990, 61-66) used two strata, each based on a government listing of adults: households that filed 1975 income tax returns, and households that were eligible for Medicaid assistance. These two lists together accounted for an estimated 96% of the population. In the selection of observations, though, these lists were sampled proportionately — 89% from the tax list and 11% from the Medicaid list — to maintain the relative proportions of these two strata in the larger population. The resulting sample can thus be claimed to represent the adult population of North Carolina. Representativeness in this case means providing the basis for accurate descriptive statistics of the entire population (e.g., average income, education, etc.).

Demographic samples are representative to the extent that they reflect the relative proportions of strata in a population. This notion of representativeness has been developed within sociological research, where researchers aim to determine descriptive statistics that characterize the overall population (such as the population mean and population standard deviation). Any single statistic that characterizes an entire population crucially depends on a proportional sampling of strata within the population — if a stratum which makes up a small proportion of the population is sampled heavily, then it will contribute an unrepresentative weight to summary descriptive statistics.

Language corpora require a different notion of representativeness, making proportional sampling inappropriate in this case. A proportional language corpus would have to be demographically organized (as discussed at the beginning of Section 3.2), because we have no *a priori* way to determine the relative proportions of different registers in a language. In fact, a simple demographically based sample of language use would be proportional by definition — the resulting corpus would contain the registers that people typically use in the actual proportions in which they are used. A corpus with this design might contain roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing. (Very few people ever produce published written texts, or unpublished written and spoken texts for a large audience.) Such a corpus would permit summary descriptive statistics for the entire language represented by the corpus. These kinds of generalizations, however, are typically not of interest for linguistic research. Rather, researchers require language samples that are representative in the sense that they include the full range of linguistic variation existing in a language.

In summary, there are two main problems with proportional language corpora. First, proportional samples are representative only in that they accurately reflect the relative numerical frequencies of registers in a language — they provide no representation of relative importance that is not numerical. Registers such as books, newspapers, and news broadcasts are much more influential than their relative frequencies indicate. Second, proportional corpora do not provide an adequate basis for linguistic analyses, in which

the range of linguistic features found in different text types is of primary interest. For example, it is not necessary to have a corpus to find out that 90% of the texts in a language are linguistically similar (because they are all conversations); rather, we want to analyze the linguistic characteristics of the other 10% of the texts, since they represent the large majority of the kinds of registers and linguistic distributions in a language.<sup>6</sup>

### 3.2 Sample size

There are many equations for determining sample size, based on the properties of the normal distribution and the sampling distribution of the mean (or the sampling distribution of the standard deviation). One of the most important equations states that the standard error of the mean for some variable ( $s_x$ ) is equal to the standard deviation of that variable ( $s$ ) divided by the square root of the sample size ( $\sqrt{N}$ ), i.e.:

$$s_x = \frac{s}{\sqrt{N}}$$

The standard error of the mean indicates how far a sample mean can be from the true population mean. If the sample size is greater than 30, then the distribution of sample means has a roughly normal distribution, so that 95% of the samples taken from a population will have means that fall in the interval of plus or minus 1.96 times the standard error. The smaller this interval is, the more confidence a researcher can have that she is accurately representing the population mean. As shown by the equation for the standard error, this confidence interval depends on the natural variation of the population (estimated by the sample standard deviation) and the sample size  $N$ . The influence of sample size in this equation is constant, regardless of the size of the standard deviation (i.e. the standard error is a function of one divided by the square-root of  $N$ ). To reduce the standard error (and thus narrow the confidence interval) by half, it is necessary to increase the sample size by four times. For example, if the sample standard deviation for the number of nouns in a text was 30, the sample mean score was 100, and the sample size was 9 texts, then the standard error would be equal to 10:

$$\text{Std. error} = \frac{30}{\sqrt{9}} = \frac{30}{3} = 10.$$

This value indicates that there is a 95% probability that the true population mean for the number of nouns per text falls within the range of 80.4 to 119.6 (i.e., the sample mean of 100 plus or minus 1.96 times the standard error of 10). To reduce this confidence interval by cutting the standard error in half, the sample size must be increased four times to 36 texts, i.e.:

$$\text{Std. error} = \frac{30}{\sqrt{36}} = \frac{30}{6} = 5.$$

---

<sup>6</sup>A proportional corpus would be useful for assessments that a word or syntactic construction is 'common' or 'rare' (as in lexicographic applications). Unfortunately, most rare words would not appear at all in a proportional (i.e., primarily conversational) corpus, making the database ill-suited for lexicographic research.

Similarly, if the original sample was 25 texts, then we would need to increase the sample to 100 texts in order to cut the standard error in half, i.e.:

$$\text{Std. error} = \frac{30}{\sqrt{25}} = \frac{30}{5} = 6.$$

$$\text{Std. error} = \frac{30}{\sqrt{100}} = \frac{30}{10} = 3.$$

Unfortunately there are certain difficulties in using the equation for the standard error to determine the required sample size of a corpus. In particular, it is necessary to address three problems:

1. The sample size  $N$  depends on a prior determination of the tolerable confidence interval required for the corpus. That is, there needs to be an *a priori* estimate of the amount of uncertainty that can be tolerated in typical analyses based on the corpus.
2. The equation depends on the sample standard deviation, but this is the standard deviation for some particular variable. Different variables can have different standard deviations, resulting in different estimates of the required sample size.
3. The equation must be used in a circular fashion. That is, it is necessary to have selected a sample and computed the sample standard deviation before the equation can be used (and this is based on the assumption that the pilot sample is at least somewhat representative) – but the purpose of the equation is to determine the required sample size.

In Section 4, I consider the distribution of several linguistic features and address these three problems, making preliminary proposals regarding sample size.

### 3.3 A Note on Sampling Within ‘Texts’

To this point I have not yet addressed the issue of how long text samples need to be. I will consider this question in more detail in Section 4, discussing the distribution of various linguistic features within texts. Here, though, I want to point out that the preference for stratified sampling applies to sampling within texts as well as across texts. Corpus compilers have typically tried to achieve better representation of texts by simply taking more words from the texts. However, these words are certainly not selected randomly (i.e., they are sequential), and the adequacy of representation thus depends on the sample length relative to the total text length. Instead it is possible to use a stratified approach for the selection of text samples from texts. That is, especially in the case of written texts and planned spoken texts, the selection of text samples can use the typical sub-components of texts in that register as sampling strata (e.g., chapters, sections, possibly main points in a lecture or sermon). This approach will result in better representation of the overall text, regardless of the total number of words selected from each text.

## **4 Distributions of Linguistic Features; Preliminary Recommendations Concerning Sample Size**

### **4.1 Distributions Within Texts; Length of Text Samples**

In this section, I consider first the distribution of linguistic features within texts, as a basis for addressing the issue of optimal text length. Traditional sampling theory is less useful here than for the other aspects of corpus design, because individual words cannot be treated as separate observations in linguistic analyses. That is, since linguistic features commonly extend over more than one word, any random selection of words from a text would fail to represent many features and would destroy the overall structure of the text. The main issue here is thus the number of contiguous words required in text samples. The present section illustrates how this issue can be addressed through empirical investigations of the distribution of linguistic features within texts.

In Biber (1990) I approach this problem by comparing pairs of 1,000-word samples taken from single texts in the LOB and London-Lund corpora. (Text samples are 2,000 words in the LOB corpus and 5,000 words in the London-Lund corpus.) If large differences are found between the two 1,000-word samples, then we can conclude that this sample length does not adequately represent the overall linguistic characteristics of a text, and that perhaps much larger samples are required. If, on the other hand, the two 1,000-word text samples are similar linguistically, then we can conclude that relatively small samples from texts adequately represent their linguistic characteristics.

In the case of written texts (from the LOB corpus), I divided each original text in half and compared the two parts. In the case of spoken texts (from the London-Lund corpus), four 1,000-word samples were extracted from each original text, and these were then compared pairwise.

To provide a relatively broad database, ten linguistic features commonly used in variation studies were analyzed. These features were chosen from different functional and grammatical classes, since each class potentially represents a different statistical distribution across text categories (see Biber 1988). The features are: first person pronouns, third person pronouns, contractions, past tense verbs, present tense verbs, prepositions, passive constructions (combining by-passives and agentless passives), WH relative clauses, and conditional subordinate clauses. Pronouns and contractions are relatively interactive and colloquial in communicative function; nouns and prepositions are used for integrating information into texts; relative clauses and conditional subordination represent types of structural elaboration; and passives are characteristic of scientific or technical styles. These features were also chosen to represent a wide range of frequency distributions in texts, as shown in Table 3, which presents their frequencies (per 1,000 words) in a corpus of 481 spoken and written texts (taken from Biber 1988, 77-78). The ten features differ considerably in both their overall average frequency of occurrence and in their range of variation. Nouns and prepositions are extremely common; present tense markers are quite common; past tense, first person pronouns, and third person pronouns are all relatively common; contractions and passives are relatively rare; and WH relative clauses and conditional subordinators are quite rare. (In addition, these features are differentially distributed across different kinds of texts; see Biber 1988.246-269). Comparison of these ten features across the 1,000-word text pairs thus represents several of the kinds

of distributional patterns found in English.

Table 3. Descriptive statistics for frequency scores (per 1,000 words) of ten linguistic features in a corpus of 481 texts taken from 23 spoken and written text genres.

Linguistic feature	Mean	Min.	Max.	Range
nouns	181	84	298	214
prepositions	111	50	209	159
present tense	78	12	182	170
past tense	40	0	119	119
third person pronouns	30	0	124	124
first person pronouns	27	0	122	122
contractions	14	0	89	89
passives	10	0	44	44
WH relative clauses	3.5	0	26	26
conditional subordination	2.5	0	13	13

The distributions of these linguistic features were analyzed in 110 1,000-word text samples (i.e., 55 pairs of samples), taken from seven text categories: conversations, broadcasts, speeches, official documents, academic prose, general fiction, and romance fiction. These categories represent a range of communicative situations in English, differing in purpose, topic, informational focus, mode, interactiveness, formality, and production circumstances; again, the goal was to represent a broad range of frequency distributions.

Reliability coefficients were computed to assess the stability of frequency counts across the 1,000-word samples. In the case of the London-Lund corpus (the spoken texts), four 1,000-word samples were analyzed from each text, and for the LOB corpus (the written texts), two 1,000-word sub-samples were analyzed from each text.

The reliability coefficient for each feature represents the average correlation among the frequency counts of that feature (i.e., a count for each of the sub-samples). For the spoken samples, all coefficients were high. The lowest reliabilities were for passives (.74) and conditional subordination (.79), while all other features had reliability coefficients over .88. The coefficients were somewhat smaller for the written samples, in part because they were based on two instead of four sub-samples. Conditional subordination in the written texts had a low reliability coefficient (.31), while relative clauses and present tense in the written texts had moderately low reliability coefficients (.58 and .61 respectively); all other features had reliability coefficients over .80. Overall, this analysis indicates that frequency counts for common linguistic features are relatively stable across 1,000 word samples, while frequency counts for rare features (such as conditional subordination and WH relative clauses – see Table 3) are less stable and require longer text samples to be reliably represented.<sup>7</sup>

<sup>7</sup>In Biber (1990) I also assess the representativeness of 1,000-word text samples by computing difference scores for pairs of samples from each text. This analysis confirms the general picture given by the reliability coefficients while providing further details of the distribution of particular features in particular registers.

These earlier analyses can be complemented by tracking the distribution of various linguistic features across 200-word segments of texts. For example, Figure 1<sup>8</sup> shows the distribution of prepositional phrases throughout the length of five texts from Humanities Academic Prose – the figure plots the cumulative number of prepositional phrases as measured at each 200-word interval in these texts. As can be seen from this figure, prepositional phrases are distributed linearly in these texts. That is, there are approximately the same number of prepositional phrases occurring in each 200-word segment (roughly 30 per segment in three of the texts, and 25 per segment in the other two texts). (The linear nature of these distributions can be confirmed by lining up a ruler next to the plot of each text.) This figure indicates that a common feature such as prepositional phrases is extremely stable in its distribution within texts (at least Humanities Academic Prose texts) – that even across 200-word segments, all segments will contain roughly the same number of prepositional phrases.

Figure 2 illustrates a curvilinear distribution, in this case the cumulative word types (i.e., the number of different words) in five Humanities texts. In general, frequency counts of a linguistic feature will be distributed linearly (although that distribution will be more or less stable within a text – see below), while frequencies of different types of linguistic features (lexical or grammatical) will be distributed curvilinearly. That is, because many types are repeated across text segments, each subsequent segment contributes fewer new types than the preceding segment. In Figure 2, the straight line marked by triangles shows the 50% boundary of word types (the score when 50% of the words in a text are different word types). In all five of these texts, at least 50% of the words are different types in the first 200-word segment (i.e., at least half of the words are not repeated), and two of the texts have more than 50% different types in the first three segments (up to 600 words). All of the texts show a gradual decay in the number of word types, however. The most diverse text drops to roughly 780 word types per 2,000 words (39%), while the least diverse text drops to roughly 480 word types per 2,000 words (only 24%). These trends would continue in longer texts, with each subsequent segment contributing fewer new types.

These two types of distributions must be treated differently. In Figures 3-9, I plot the distributions of seven linguistic features within texts representing three registers. Three of the features are cumulative frequency counts: Figure 3 plots the frequencies of prepositional phrases, a common grammatical feature; Figure 4 plots the frequencies of relative clauses, a relatively rare grammatical feature; and Figure 5 plots the frequencies of noun-preposition sequences, a relatively common grammatical sequence. The other four figures plot the distributions of types in texts. Figures 6 and 7 plot the distribution of lexical types: word types (the number of different words) in Figure 6 and hapax legomena (once-occurring words) in Figure 7. Figures 8 and 9 plot the distribution of grammatical types: different grammatical categories or ‘tags’ in Figure 8, and different grammatical tag sequences in Figure 9. The figures thus illustrate lexical and grammatical features, with rare and common overall frequencies, having linear and curvilinear distributions.

The figures can be used to address several questions. First they present the overall distributions of these features. The stable linear distribution of prepositional phrases is further confirmed by Figure 3. In contrast, the relatively unstable distribution of relative clauses, indicated above by a relatively low reliability coefficient, is further supported

---

<sup>8</sup>Figures 1-9 are found in the appendix.

by the frequent departures from linearity in Figure 4. That is, since relative clauses are relatively rare overall, even two or three extra relatives in a 200-word segment results in an aberration. Figure 5 shows that the distribution of noun-preposition sequences is similar to that of prepositional phrases in being linear and quite stable (although less frequent overall).<sup>9</sup>

Figures 6-9 show different degrees of curvilinearity, with the grammatical and syntactic types showing sharper drop-offs than the lexical types. Grammatical tag types show the sharpest decay: most different grammatical categories occur in the first 200 words, with relatively few additional grammatical categories being added after 600 words.

Figures 3-9 also illustrate distributional differences across registers, although only three registers are considered here. For example, Figures 3 and 5 show fairly large differences between academic prose and fiction, with the former having much higher frequencies of prepositional phrases and noun-prepositional phrase sequences. The differences among registers are less clear-cut in Figure 4, but humanities academic prose texts consistently have more frequent relative clauses than either technical academic prose or fiction.

Each register is plotted twice in these figures: the ‘average’ scores and the ‘10-text’ scores. Average scores present the average value for ten texts from that register for the segment in question. (For example, Figure 3 shows that humanities texts have on average 130 prepositions in the first 1,000 words of text.) In contrast, the ‘10-text’ scores are composite scores, with each 200-word segment coming from a different text. Thus, the score for 400 words represents the cumulative totals for the first 200 words from two texts, the score for 600 words sums the first 200-word totals from three texts, etc.

In the case of stable, linear distributions, there is very little difference between the average and 10-text scores. In fact, Figures 3 and 5 show a remarkable coincidence of average and 10-text values; a single distribution is found, within a register, regardless of whether subsequent 200-word segments are taken from the same text or from different texts. Figure 4 shows greater differences for relative clauses (a relatively rare and less stable feature). Here averaging over ten texts smooths out most aberrations from linearity, while the 10-text values show considerable departures from linearity.

In contrast, there are striking differences between the average and 10-text distributions for the curvilinear features (Figures 6-9). In these cases, the 10-text scores are consistently higher than the corresponding average score from the same register. In the case of word types, the 10-text scores for all three registers are higher than the average scores of the registers. The difference is particularly striking with respect to technical academic prose — after 2,000 words of text, 10-text technical prose has the second highest word type score (approximately 740, or 37%), while on average technical prose texts have the lowest word type score (approximately 570, or 28%). This shows that there is a high degree of lexical repetition within technical prose texts, but there is a high degree of lexical diversity across technical texts. The distribution of hapax legomena, shown in Figure 7, parallels that of word types — again, all three 10-text scores are higher than the average scores; 10-text humanities prose shows the highest score; and the average technical prose score is by far the lowest. These distributions reflect the considerable

---

<sup>9</sup>These are primarily prepositional phrases functioning as noun modifiers, as opposed to prepositional phrases with adverbial functions.

lexical diversity found across humanities texts, and the relatively little lexical diversity within individual technical texts.

There is more similarity between the 10-text and average scores with respect to the distribution of grammatical types (Figures 8 and 9), although for each register the 10-text score is higher than the corresponding average score. Interestingly, these figures show that technical prose has the least grammatical diversity as well as the least lexical diversity.

In summary, the analyses presented in this section indicate the following:

1. Common linear linguistic features are distributed in a quite stable fashion within texts and can thus be reliably represented by relatively short text segments.
2. Rare linguistic features show much more distributional variation within texts and thus require longer text samples for reliable representation.
3. Features distributed in a curvilinear fashion, i.e., different feature types, are relatively stable across subsequent text segments, but occurrences of new types decrease throughout the course of a text. The frequency of new types is consistently higher in cross-text samples than in single-text samples. These patterns were shown to hold for relatively short text segments (2,000 words total) and for cross-text samples taken from a single register; the patterns should be even stronger for longer text segments and for cross-register samples. These findings support the preference for stratified sampling — more diversity among the texts included in a corpus will translate into a broader representation of linguistic feature types.

With regard to the issue of text length, the thrust of the present section is simply that text samples should be long enough to reliably represent the distributions of linguistic features. For linearly-distributed features, the required length depends on the overall stability of the feature. For curvilinear features, an arbitrary cut-off must be specified marking an ‘adequate’ representation, for example, when subsequent text segments contribute less than 10% additional new types. Given a finite effort invested in developing a corpus, broader linguistic representation can be achieved by focusing on diversity across texts and text types rather than by focusing on longer samples from within texts.

Specific proposals on text length require further investigations of the kind presented here, focusing especially on the distributions of less stable features, to determine the text length required for stability, and on the distributions of other kinds of features (e.g., discourse and information-packaging features).

## 4.2 Distributions Across Texts; Number of Texts

A second major statistical issue in building a text corpus concerns the sampling of texts: how linguistic features are distributed across texts and across registers, and how many texts must be collected for the total corpus and for each register to represent those distributions.

**4.2.1 Previous research on linguistic variation within and across registers** Although registers are defined by reference to situational characteristics, they can be analyzed linguistically, and there are important linguistic differences among them; at the same time, some registers also have relatively large ranges of linguistic variation internally (see Biber 1988, Chapters 7 and 8). For this reason, the linguistic characterization of a register should include both its central tendency and its range of variation. In fact, some registers are similar in their central tendencies but differ markedly in their ranges of variation (e.g., science fiction versus general fiction, and official documents versus academic prose, where the first register of the pair has a more restricted range of variation). In Biber (1988, 170-198), I describe the linguistic variation within registers, including the linguistic relations among various sub-registers.

The number of texts required in a corpus, and to represent particular registers, relates directly to the extent of internal variation. In Biber (1990), I analyze the stability of feature counts across texts from a register by comparing the mean frequency counts for 10-text sub-samples taken from particular registers. Five registers were analyzed: conversations, public speeches, press reportage, academic prose, and general fiction. Three 10-text samples were extracted from each of these registers, and the mean frequency counts of six linguistic features were compared across the samples (first person pronouns, third person pronouns, past tense, nouns, prepositions, passives). The reliability analysis of these mean frequencies across the three 10-text samples showed an extremely high degree of stability for all six linguistic features (all coefficients greater than .95). These coefficients show that the mean scores of the 10-text samples are very highly correlated; that is, the central linguistic tendencies of these registers with respect to these linguistic features are quite stable, even as measured by 10-text samples. However, there are two important issues not considered by this analysis. First, the six linguistic features considered were all relatively common; rare features such as WH relative clauses or conditional subordination might show much lower reliabilities. Second, this analysis addressed how many texts were needed to reliably represent mean scores, but did not address the representation of linguistic diversity in registers.<sup>10</sup>

**4.2.2 Total linguistic variation in a corpus; total sample size for a corpus.** In Section 3, I discussed how the required sample size is related to the standard error ( $s_x$ ) by the equation:

$$4.1 \quad s_x = \frac{s}{\sqrt{N}}$$

The actual computation of sample size depends on a specification of the tolerable error ( $te$ ):

$$4.2 \quad t_e = t * s_x$$

Equation 4.2 states that the tolerable error is equal to the standard error times the  $t$ -value. Given a sample size greater than 30 (which permits the assumption of a normal distribution), a researcher can know with 95% confidence that the mean score of a sample will fall in the interval of the true population mean plus-or-minus the tolerable error.

---

<sup>10</sup>Actually this latter question was addressed by computing difference scores, for the mean, standard deviation, and range, across the 10-text samples.

Equation 4.2 can be manipulated to provide a second equation for computing the standard error, i.e.,  $s_x = \frac{t_e}{t}$ . If the ratio  $t_e/t$  is substituted for  $s_x$  in Equation 4.1, and the equation is then solved for  $N$ , we get a direct computation of the required sample size for a corpus:

$$4.3 \quad N = \frac{s^2}{\left(\frac{t_e}{t}\right)^2}$$

where  $N$  is the computed sample size,  $s$  is the estimated standard deviation for the population,  $t_e$  is the tolerable error (equal to 1/2 of the desired confidence interval), and  $t$  is the  $t$ -value for the desired probability level.

I note in Section 3 that there are problems in the application of Equation 4.3. In one sense, the equation simply shifts the burden of responsibility, from estimating the unknown quantity for required sample size to estimating the unknown quantities for the tolerable error and population standard deviation. That is, in order to use the equation, there needs to be a prior estimate of the tolerable error or confidence interval permitted in the corpus and a prior estimate of the standard deviation of variables in the population as a whole.

The tolerable error depends on the precision required of population estimates based on the corpus sample. For example, say that we want to know how many nouns on average occur in conversation texts. The confidence interval is the window within which we can be 95% certain that the true population mean falls. For example, if the sample mean for nouns in conversations was 120, and we needed to estimate the true population mean of nouns with a precision of plus or minus 2, then the confidence interval would be 4, extending from 118 to 122. The tolerable error is simply one side (or one-half) of the confidence interval. The problem here is that it is difficult to provide an *a priori* estimate of the required precision of the analyses that will be based on a corpus.

Similar problems arise with the estimation of standard deviations. In this case, it is not possible to estimate the standard deviation of a variable in a corpus without already having a representative sample of texts. Here, as in many aspects of corpus design, work must proceed in a circular fashion, with empirical investigations based on pilot corpora informing the design process. The problem for initial corpus design, however, is to provide an initial estimate of standard deviation.

A final problem is that standard deviations must be estimated for particular variables, but in the case of corpus linguistics, there are numerous linguistic variables of interest. Choosing different variables, with different standard deviations, will result in different estimates of required sample size.

In the present section, I use the analyses in Biber (1988, 77-78, 246-269) to address the first two of these problems. That study is based on a relatively large and wide-ranging corpus of English texts: 481 texts taken from 23 spoken and written registers. Statistical analyses of this corpus can thus be used to provide initial estimates for both the tolerable error and the population standard deviation.

In the design of a text corpus, tolerable error cannot be stated in absolute terms because the magnitude of frequency counts varies considerably across features (as was shown in Section 3). For example, a tolerable error of plus or minus 5 might work well for common features such as nouns, which have an overall mean of 180.5 per

1,000 words in the pilot corpus, but it would be unacceptable for rare features such as conditional subordinate clauses, which have an overall mean of only 2.5 in the corpus (so that a tolerable error of 5 would translate into a confidence interval of –2.5 to 7.5, and a text could have three times the average number of conditional clauses and still be within the confidence interval). Instead I propose here computing a separate estimate of the tolerable error for each linguistic feature, based on the magnitude of the mean score for the feature; for illustration, I will specify the tolerable error as plus-or-minus 5% of the mean score (for a total confidence interval of 10% of the mean score). Table 4 presents the mean score and standard deviation of seven linguistic features in the pilot corpus, together with the computed tolerable error for each feature. It can be seen that the tolerable error ranges from 9.03 for nouns (which have a mean of 180.5) to 0.13 for conditional clauses (which have a mean of only 2.5).

Table 4. Estimates of required sample sizes (number of texts) for the total corpus.

	Mean score in pilot corpus	std. dev. in pilot corpus	tolerable error	required <i>N</i>
Nouns	180.5	35.6	9.03	59.8
Prepositions	110.5	25.4	5.53	81.2
Present tense	77.7	34.3	3.89	299.4
Past tense	40.1	30.4	2.01	883.1
Passives	9.6	6.6	0.48	726.3
WH Relative cls.	3.5	1.9	0.18	452.8
Conditional cls.	2.5	2.2	0.13	1,190.0

Given the tolerable errors and estimated standard deviations listed in Table 4, required sample size (i.e., the total number of texts to be included in the corpus) can be computed directly using Equation 4.3. Table 4 shows very large differences in required sample size across these linguistic features. These differences are a function of the size of the standard deviation relative to the mean for a particular feature. If the standard deviation is many times smaller than the mean, as in the case of common features such as nouns and prepositions, the required sample size is quite small. If, on the other hand, the standard deviation approaches the mean in magnitude, as in the case of rare features such as WH relative clauses and conditional clauses, the required sample size becomes quite large. Past tense markers are interesting in that they are relatively common (mean of 40.1) yet have a relatively large standard deviation (30.4) and thus require a relatively large sample of texts for representation (883). Overall the most conservative approach in designing a corpus would be to use the most widely varying feature (proportional to its mean – in this case conditional clauses) to set the total sample size.

**4.2.3 Linguistic variation within registers; number of texts needed to represent registers.** The remaining issue concerns the required sample size for each register. Although most books on sample design simply recommend proportional sampling for stratified designs (see Section 3), a few books discuss the need for non-proportional

stratified sampling in certain instances; these books differ, however, on the method for determining the recommended sample sizes for subgroups. For example, Sudman (1976, 110-111) states that non-proportional stratified sampling should be used when the subgroups themselves are of primary interest (as in the case of a text corpus), and that the sample sizes of the subgroups should be equal in this case (to minimize the standard error of the difference). This procedure is appropriate when the variances of the subgroups are roughly the same. In contrast, Kalton (1983, 24-25) recommends using the sub-group standard deviations to determine their relative sample sizes. This procedure is more appropriate for corpus design, since the standard deviations of linguistic features vary considerably from one register to the next.

Although I do not make specific recommendations for register sample size here, I illustrate this approach in Table 5, considering the relative variances of seven linguistic features (the same as in Table 4) across three registers: conversations, general fiction, and academic prose. As above, the data are taken from Biber (1988, 246-269).

Table 5 presents the mean score, standard deviation, and the ratio of standard deviation to mean score, for these seven linguistic features in the three registers. The ratio represents the normalized variance of each of these features within each register – the extent of internal variation relative to the magnitude of the mean score. The raw standard deviation is not appropriate here (similar to Table 4) because the mean scores of these features varies to such a large extent.

Table 5 shows that the normalized standard deviation varies considerably across features within a register. Within conversations, for example, the counts for nouns, prepositions, and present tense all show relatively small normalized variances, while passives, WH relative clauses, and conditional clauses all show normalized variances at or above 50%. As shown earlier, features with lower overall frequencies tend to have considerably higher normalized variances.

There are also large differences across the registers. For example, past tense has a normalized variance of 46% in conversations and only 18% in general fiction, but it shows a normalized variance of 96% in academic prose. Conditional subordination also shows large differences across these three registers: it has a normalized variance of 54% in conversations, 73% in general fiction, and 100% in academic prose.

In order to determine the sample size for each register, it is necessary to compute a single measure of the variance within each register. This measure is then used to allot a proportionally larger sample to registers with greater variances. (This should not be confused with a proportional representation of the registers.) A certain minimum number of texts should be allotted for each register (e.g., at least 20 texts per register), and then the remaining texts in the corpus can be divided proportionally depending on the relative variance within registers.

To illustrate, consider Table 5 again. This table lists an average normalized deviation for each register, which represents an overall deviation score computed by averaging the normalized standard deviations of the seven linguistic features. Conversations and general fiction both have relatively similar overall deviations (37% and 39% correspondingly) while academic prose has a somewhat higher overall deviation (49%). To follow through with this example, assume that there were to be a total of 200 texts in a corpus, taken from these three registers. Each register would be allotted a minimum of 20 texts, leaving 140 texts to be divided proportionally among the three registers. To determine

Table 5. Relative variation within selected registers.

Conversations: Average normalized deviation = .37

	Mean score in pilot corpus	std. dev. in pilot corpus	ratio of std dev / mean (normalized deviation)
Nouns	137.4	15.6	.11
Prepositions	85.0	12.4	.15
Present tense	128.4	22.2	.17
Past tense	37.4	17.3	.46
Passives	4.2	2.1	.50
WH Relative cls.	1.4	0.9	.64
Conditional cls.	3.9	2.1	.54

General Fiction: Average normalized deviation = .39

	Mean score in pilot corpus	std. dev. in pilot corpus	ratio of std dev / mean (normalized deviation)
Nouns	160.7	25.7	.16
Prepositions	92.8	15.8	.17
Present tense	53.4	18.8	.35
Past tense	85.6	15.7	.18
Passives	5.7	3.2	.56
WH Relative cls.	1.9	1.1	.58
Conditional cls.	2.6	1.9	.73

Academic Prose: Average normalized deviation = .49

	Mean score in pilot corpus	std. dev. in pilot corpus	ratio of std dev / mean (normalized deviation)
Nouns	188.1	24.0	.13
Prepositions	139.5	16.7	.12
Present tense	63.7	23.1	.36
Past tense	21.9	21.1	.96
Passives	17.0	7.4	.44
WH Relative cls.	4.6	1.9	.41
Conditional cls.	2.1	2.1	1.00

the relative sample size of the registers, one would solve the following equation based upon their relative overall deviations:

$$\begin{aligned}.37x + .39x + .49x &= 140 \\1.25x &= 140 \\x &= 112\end{aligned}$$

and thus the sample sizes would be:

conversation:  $.37 * 112 = 41$

general fiction:  $.39 * 112 = 44$

academic prose:  $.49 * 112 = 55$

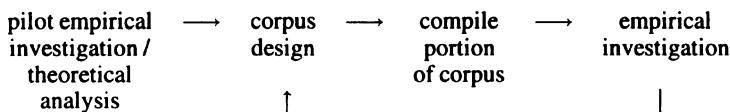
$$\begin{aligned}\text{total allocated texts} &= (41 + 20 \text{ for conversation}) + \\&\quad (44 + 20 \text{ for general fiction}) + \\&\quad (55 + 20 \text{ for academic prose}) \\&= 200 \text{ texts}\end{aligned}$$

To compute the actual values for register sample sizes, it is necessary to analyze the full range of linguistic features in all registers, computing a single average deviation score for each register. This could be done by averaging across the normalized variances of all linguistic features, as illustrated here. An alternative approach would be to use the normalized variances of the linguistic dimensions identified in Biber (1988). This latter approach would have a more solid theoretical foundation, in that the dimensions represent basic parameters of variation among registers, each based on an important co-occurrence pattern among linguistic features. In contrast, the approach illustrated in this section depends on the pooled influence of linguistic features in isolation, and thus relatively aberrant distributions can have a relatively strong influence on the final outcome. In addition, use of the dimensions enables consideration of the distributions with respect to particular functional parameters, so that some dimensions can be given more weight than others. In contrast, there is no motivated way for distinguishing among the range of individual features on functional grounds.

It is beyond the scope of this paper to illustrate the use of dimension scores for the linguistic characterization of registers (since I would first need to explain the theoretical and methodological bases of the dimensions). The same basic approach as illustrated in this section would be used, however. The major difference involves the analysis of deviation along basic dimensions of linguistic variation rather than with respect to numerous linguistic features in isolation.

## 5 Conclusion: Beginning

I have tried to develop here a set of principles for achieving 'representativeness' in corpus design. I have offered specific recommendations regarding some aspects of corpus design, and illustrations elsewhere (regarding issues for which final recommendations could not be developed in a paper of this scope). The bottom-line in corpus design, however, is that the parameters of a fully representative corpus cannot be determined at the outset. Rather, corpus work proceeds in a cyclical fashion that can be schematically represented as follows:



Theoretical research should always precede the initial corpus design and actual compilation of texts. Certain kinds of research can be well advanced prior to any empirical investigations: identifying the situational parameters that distinguish among texts in a speech community, and identifying the range of important linguistic features that will be analyzed in the corpus. Other design issues, though, depend on a pilot corpus of texts for preliminary investigations. Present-day researchers on English language corpora are extremely fortunate in that they have corpora such as the Brown, LOB, and London-Lund corpus for pilot investigations, providing a solid empirical foundation for initial corpus design. The compilers of those corpora had no such pilot corpus to guide their designs. Similar situations exist for current projects designing corpora to represent non-western languages. For example, a recently completed corpus of Somali required extensive fieldwork to guide the initial design (see Biber and Hared 1992). Thus the initial design of a corpus will be more-or-less advanced depending on the availability of previous research and corpora.

Regardless of the initial design, the compilation of a representative corpus should proceed in a cyclical fashion: a pilot corpus should be compiled first, representing a relatively broad range of variation but also representing depth in some registers and texts. Grammatical tagging should be carried out on these texts, as a basis for empirical investigations. Then empirical research should be carried out on this pilot corpus to confirm or modify the various design parameters. Parts of this cycle could be carried out in an almost continuous fashion, with new texts being analyzed as they become available, but there should also be discrete stages of extensive empirical investigation and revision of the corpus design.

Finally, it should be noted that various multivariate techniques could be profitably used for these empirical investigations. In this paper, I have restricted myself to univariate techniques, and to simple descriptive statistics. My own research, though, suggests the usefulness of two multivariate techniques for the analysis of linguistic variation in computerized corpora: factor analysis and cluster analysis. Factor analysis can be used in either an exploratory fashion (e.g., Biber 1988) or for theory-based 'confirmatory' analyses (e.g., Biber 1992). Both of these would be appropriate for corpus design work, especially for the analysis of the range and types of variation within a corpus and within registers. Such analyses would indicate whether the different parameters of variation were equally well represented and would provide a basis for decisions on sample size. Cluster analysis has been used to identify 'text types' in English – text categories defined in strictly linguistic terms (Biber 1989). Text types cannot be identified on *a priori* grounds; rather they represent the groupings of texts in a corpus that are similar in their linguistic characterizations, regardless of their register categories. Ideally a corpus would represent both the range of registers and the range of text types in a language, and thus research on variation within and across both kinds of text categories is needed.<sup>11</sup>

<sup>11</sup>For example, one of the most marked text types identified in Biber (1989) consists of texts in which the addressee is producing an on-line reportage of events in progress. Linguistically, this text type is marked in being extremely situated in reference (many time and place adverbials and a present time orientation).

In sum, the design of a representative corpus is not truly finalized until the corpus is completed, and analyses of the parameters of variation are required throughout the process of corpus development in order to fine-tune the representativeness of the resulting collection of texts.

## References

- [1] Biber, Douglas. 1986. Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62.384-414.
- [2] Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- [3] Biber, Douglas. 1989. A typology of English texts. *Linguistics*, 27.3-43.
- [4] Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5.
- [5] Biber, Douglas. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15.133-163.
- [6] Biber, Douglas. 1993a. An analytical framework for register studies, *Sociolinguistic perspectives on register* ed. by D. Biber, and E. Finegan, New York: Oxford University Press. (in press).
- [7] Biber, Douglas. 1993b. Register variation and corpus design. To appear in *Computational Linguistics*.
- [8] Biber, Douglas, and Mohamed Hared. 1992. Dimensions of register variation in Somali. *Language Variation and Change* 4.41-75.
- [9] Brown, Penelope, and Colin Fraser. 1979. Speech as a marker of situation. *Social markers in speech*, ed. by Klaus R. Scherer and Howard Giles, 33-62. Cambridge: Cambridge University Press.
- [10] Duranti, Alessandro. 1985. Sociocultural dimensions of discourse. *Handbook of discourse analysis* (Vol. 1), ed. by Teun van Dijk, 193-230. New York: Academic Press.
- [11] Francis, W. Nelson, and Henry Kučera. 1964/1979. Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Department of Linguistics, Brown University.
- [12] Halliday, Michael A.K., and Ruqaiya Hasan. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

---

Unfortunately, there are only 7 such texts in the combined London-Lund and LOB corpora, indicating that this text type is under-represented and needs to be targeted in future corpus development.

- [13] Henry, Gary T. 1990. *Practical sampling*. Newbury Park, CA: Sage.
- [14] Hymes, Dell H. 1974. *Foundations in sociolinguistics*. Philadelphia: University of Pennsylvania Press.
- [15] Johansson, Stig, Geoffrey N. Leech, and Helen Goodluck. 1978. Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers. Department of English, University of Oslo.
- [16] Kalton, Graham. 1983. *Introduction to survey sampling*. Newbury Park, CA: Sage.
- [17] Sudman, Seymour. 1976. *Applied sampling*. New York: Academic Press.
- [18] Svartvik, Jan, and Randolph Quirk (eds.). 1980. *A corpus of English conversation*. Lund: C.W.K. Gleerup.
- [19] Williams, Bill. 1978. *A sampler on sampling*. New York: John Wiley and Sons.

## Appendix

Figure 1: Distribution of Prepositions  
in Five Humanities Texts

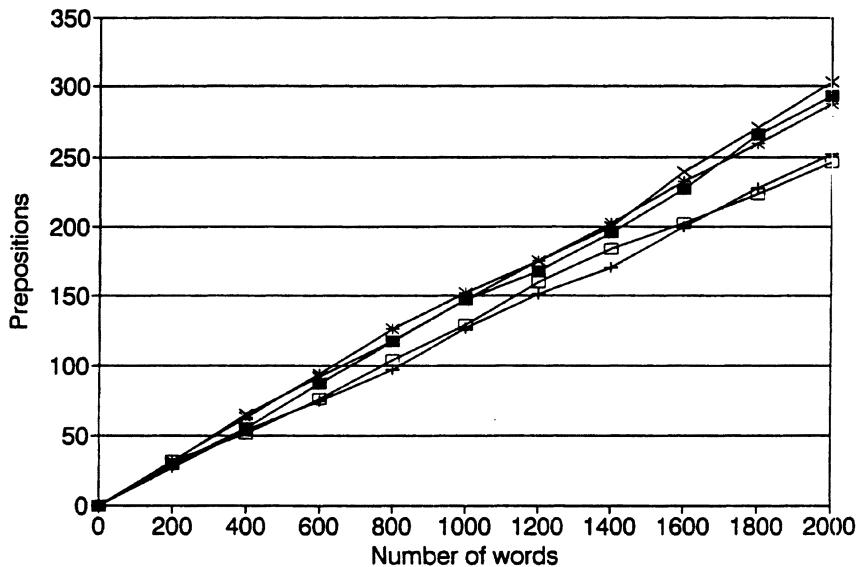
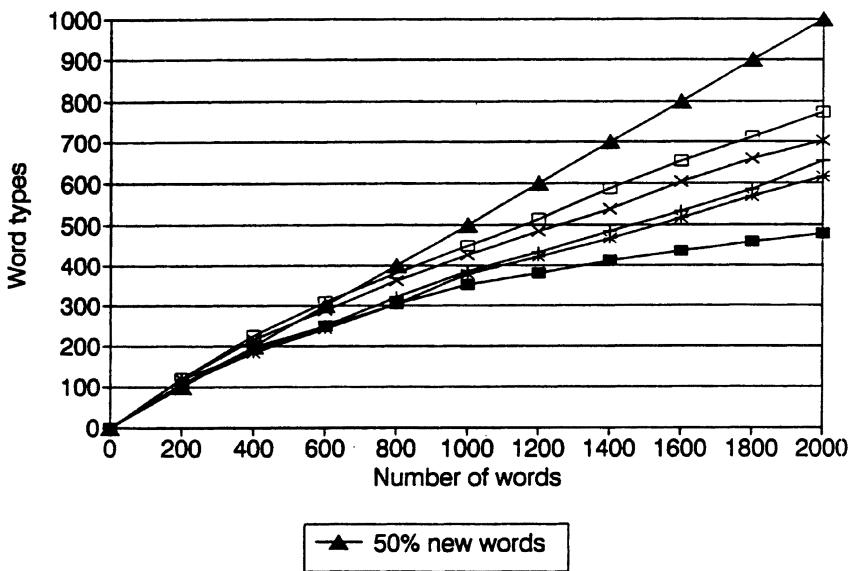
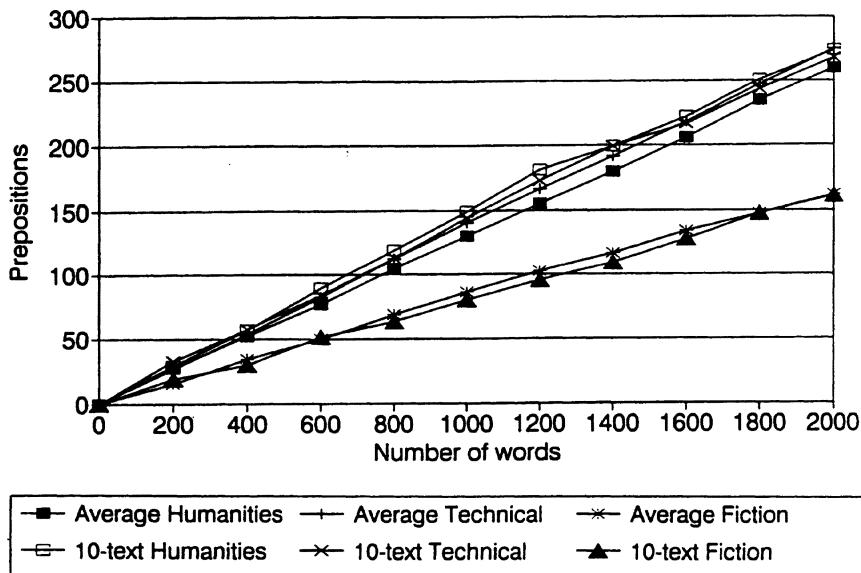


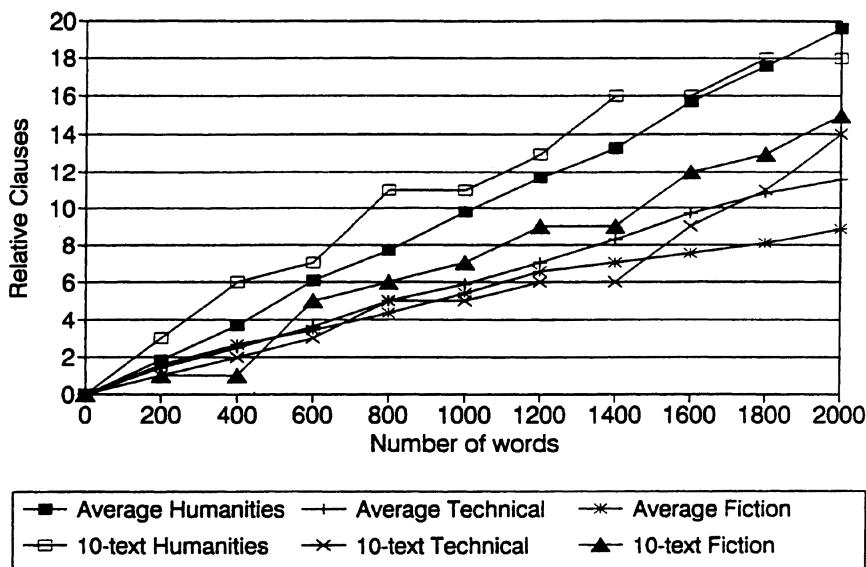
Figure 2: Distribution of Word Types  
in Five Humanities Texts



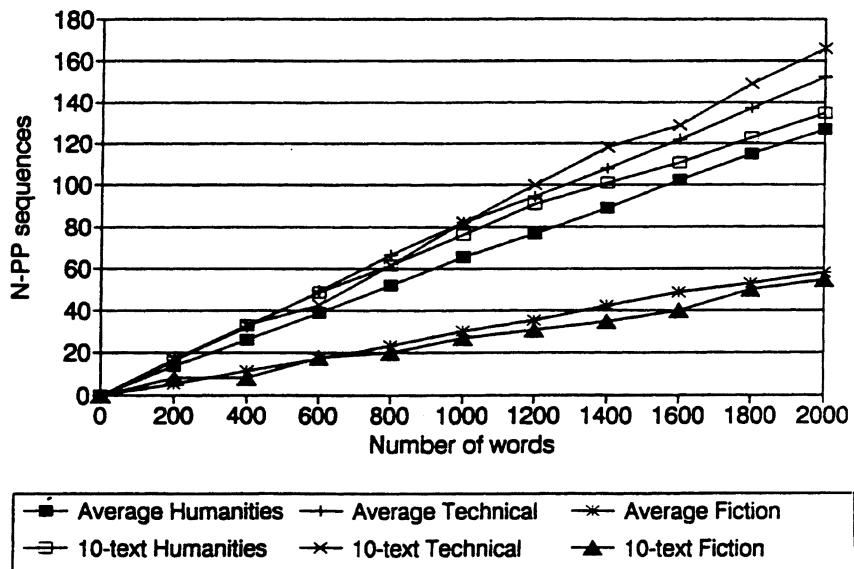
**Figure 3: Distribution of Prepositions  
in Texts from Three Registers**



**Figure 4: Distribution of Relative Cls.  
in Texts from Three Registers**



**Fig. 5: Distribution of N-PP Sequences  
in Texts from Three Registers**



**Figure 6: Distribution of Word Types  
in Texts from Three Registers**

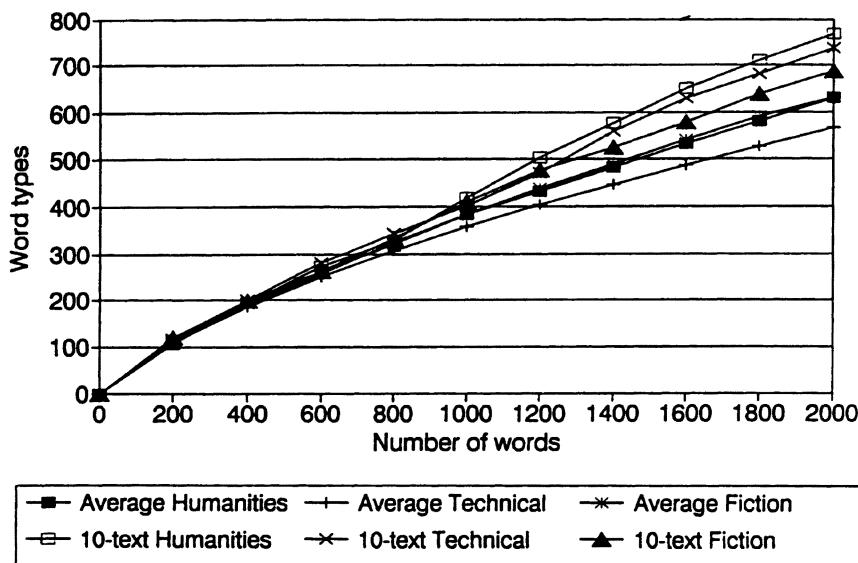


Fig. 7: Distribution of Hapax Legomena  
in Texts from Three Registers

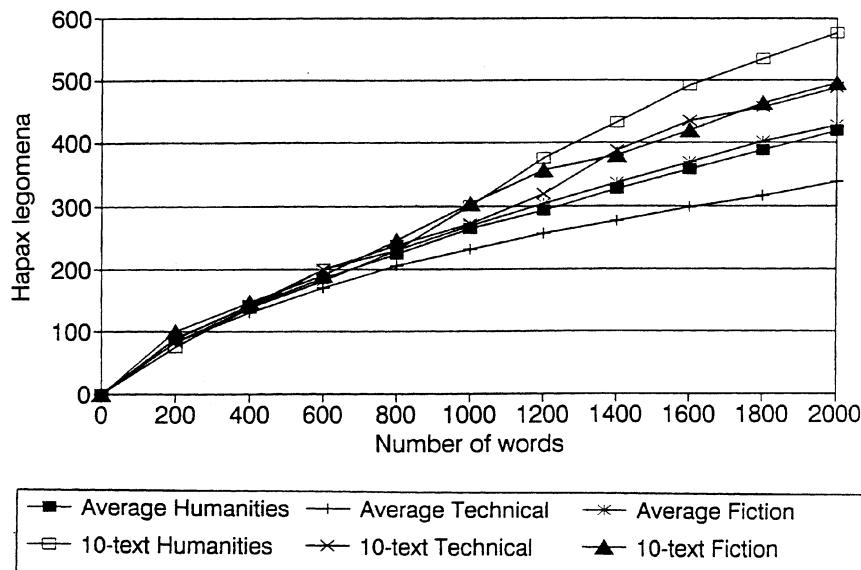


Figure 8: Distribution of Grammatical Tag Types in Texts from Three Registers

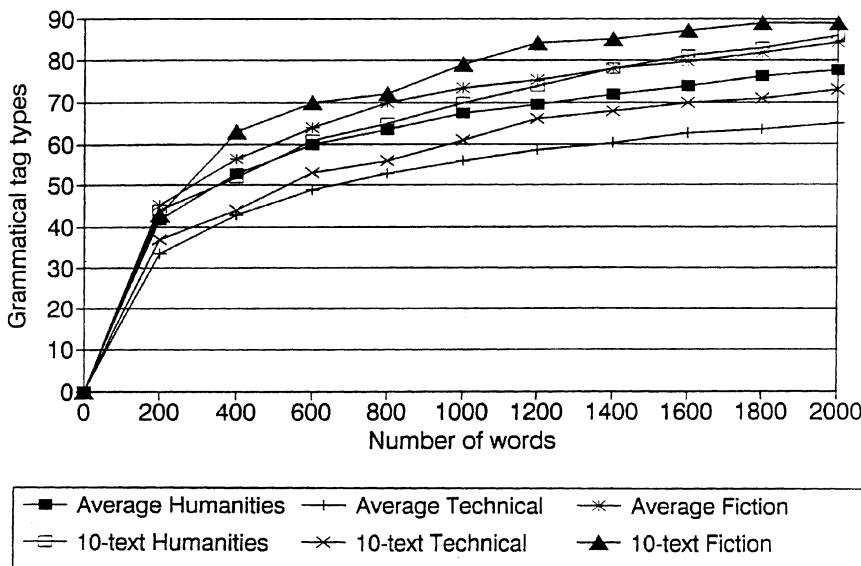
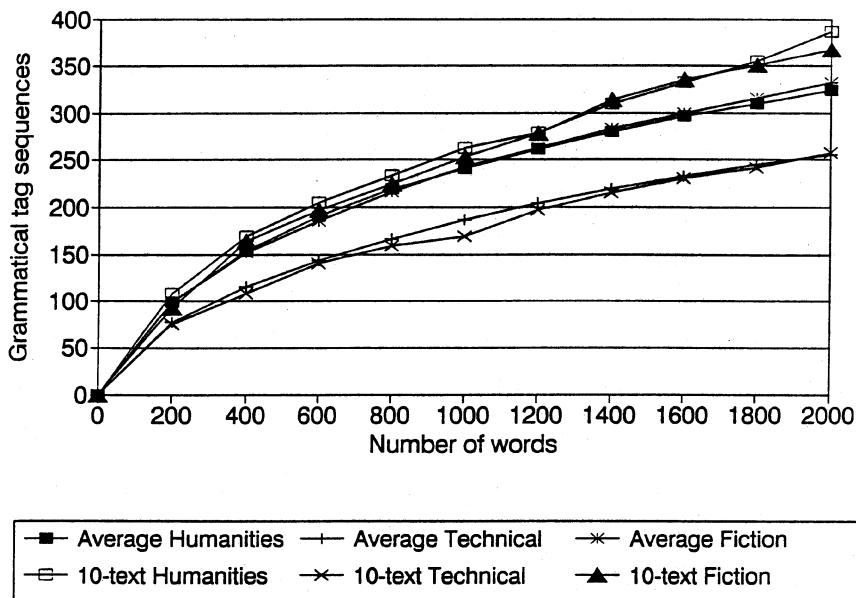


Figure 9: Distribution of Grammatical Tag Sequences in Texts from 5 Registers



# The Text Encoding Initiative

C. M. Sperberg-McQueen  
University of Illinois at Chicago  
*e-mail: u35395@uicvm.bitnet*

## Abstract

This paper describes the goals and work of the Text Encoding Initiative, an international cooperative project to develop and disseminate guidelines for the encoding and interchange of electronic text for research purposes. It begins by outlining some basic problems which arise in the attempt to represent textual material in computers, and some problems which arise in the attempt to encourage the sharing and reuse of electronic textual resources. These problems provide the necessary background for a brief review of the origins and organization of the Text Encoding Initiative itself. Next, the paper describes the rationale for the decision of the TEI to use the Standard Generalized Markup Language (SGML) as the basis for its work. Finally, the work accomplished by the TEI is described in general terms, and some attempt made to clarify what the project has and has not accomplished.

## 1 Introduction

This paper describes the goals and work of the Text Encoding Initiative, an international cooperative project to develop and disseminate guidelines for the encoding and interchange of electronic text for research purposes, with which Don Walker was associated from its beginning in 1987 until the end of his life. In the simplest possible terms, the Text Encoding Initiative (henceforth: TEI) is an attempt to find better ways to put texts into computers for the purposes of doing research which uses those texts.

The paper will discuss first some basic problems involved in that process, then some practical aspects of the reuse and reusability of textual resources. With the context thus clarified, the origins and organization of the TEI itself can then be described briefly, along with the reasons behind the decision of the TEI to use the Standard Generalized Markup Language (SGML) as the basis for its work. Finally, the work accomplished by the TEI can be described in general terms, and some attempt made to clarify what the project has and has not accomplished.

## 2 Electronic Text as a Resource

In the introductory paragraph of this paper the TEI was described as an international cooperative effort to find better ways of “putting texts into computers”. The first problem encountered when one tries to set about this task is that, in a literal sense, it cannot be

done. Texts *cannot* be placed inside computers. Neither can numbers — if only for the pedantic but simple reason that numbers are abstract mathematical objects, and texts are abstract linguistic, literary, aesthetic, referential, historical, and cultural objects, while computers are physical objects controlled by complex electronic circuitry. Abstract objects cannot be “put into” physical objects. The best one can do is to make the physical object mimic the salient features of the abstract object, and to manipulate this *physical representation* of the abstract object.

The value in this admittedly pedantic quibble is that it forces us to face squarely the critical fact that our problem is thus one of *mimesis* (or to put it into computational terms, one of finding a suitable *representation* for the data). Instead of a simple mechanical or quasi-mechanical process which can be carried out without reflection, the representation of texts in electronic form involves the same complications and limitations inherent in any act of representation. Representations never reproduce all aspects of their objects with perfect fidelity; they invariably omit some aspect or other of the object represented, and by this omission distort it. Designing a method for representing some object by means of some other object therefore ineluctably requires the designer not simply to decide what is salient and must be included, but equally what is expendable and may be tossed off the sled. It is no wonder, then, that systematic schemes for the representation of whole classes of objects reflect the biases, preconceptions, and preoccupations of their designers.

And yet for all their flaws, representations are absolutely essential to any intellectual work at all, because they are essential to understanding. Because they are selective reproductions of what is thought salient about some object, representations serve to reify our understanding of the object represented, and allow us to test that understanding and compare it with different views of the object — themselves reified by different representations.

These issues are familiar, in a restricted form, to any computer programmer who has had to consider whether to represent a numeric quantity as a short integer, a long integer, or a real number at single or double precision; they are much less widely familiar when it comes to the representation of textual data in electronic form, even though textual data are intellectually much more complex and much less well defined than integers and real numbers — perhaps in part *because* text is less well defined.

If, as Niklaus Wirth has put it, “programs = algorithms + data structures,” then a suitable method of representing textual data might be expected to represent a significant step forward in computational work with language and literature. Such a representation should make it easier to use computers to work with texts, and thus contribute to the success of textual research and indirectly to the understanding of texts and of textual information.

If one asks oneself about the nature of a suitable representation for texts in electronic form, what it would mean to “represent a text” in a machine, one discovers a second advantage of the pedantic quibble with which this paper began. For, being forced to pose this question in terms of representations, one is equally forced to recognize that — since representations are typically utilitarian in character — the answer will inevitably be “it depends; suitable for what?” Before defining the qualities of a “suitable representation,” one must specify what use is to be made of it. One is thus led to ask what it is that those interested in text in electronic form want to *do* with it.

A first simple answer is that we want to use it in the normal manner. Since it is text and we are readers, we will want to read it. Users will want to disseminate it to friends, colleagues, or the public across the network. As researchers, they will want to study it: literary scholars will want to study its themes, images, style, narrative structure, vocabulary, and diction; linguists will want to study its lexicon, morphology, parts of speech, syntax, or discourse structure. Textual critics will want to edit it, to study the variants in different manuscripts or early editions of the work, collate the various versions, and annotate it. Even those who work most intensively with computers will probably want to print the text out, nicely formatted, on paper. As time passes, the chances are good that over time people will want to link the text to related material, be it other versions of the same text, commentary, graphics or illustrations, images of manuscripts, or yet other materials, either locally in a network environment.

Equally important, we will want to *reuse* it. The costs of getting material into an acceptable electronic form are high enough to make reuse of data an important goal in virtually every computational field, from the natural sciences to the social sciences to the humanities. In the humanities, this fact is reflected in the increasing numbers of projects whose aim is to create generally usable bodies of electronic textual material intended for use by others; in computational linguistics, it is reflected in the growth of projects to develop standard reference corpora for use in all areas of natural language processing, as well as in efforts to create “opportunistic corpora” gathering together as much textual material as can be obtained.<sup>1</sup>

Third, because many of those interested in electronic text are researchers, it is a safe prediction that they will eventually want to do things with this electronic text which no one has yet invented or imagined. It is in the nature of research that not only the answers to the questions, but frequently the questions themselves, are not known at the outset of a project.

In other words, there is no satisfactory answer to the question what we want to do with texts, once we put them into electronic form. In the long run, we want to do *everything*. This is not a wholly vacuous answer to the question; it does have the consequence that we want a representation which, as far as possible, does not constrain what we can do with the text. Anything we can do with the text, we would like to be able to do with the representation. It also has the consequence that we should resist the temptation to design the electronic representation of text with any single application in mind. Since any given application for the electronic text is only one among many, there is not much point in designing characteristics into our data representation which make sense only in one application: a more general representation will make better sense in

---

<sup>1</sup> Among humanities projects, one might mention the Brown University Women Writers Project, which is creating a corpus of women's writing in English from 1330 to 1830; the Nietzsche Nachlaßproject now at Dartmouth; the Leiden Armenian Database, collecting primarily medieval Armenian texts; the Global Jewish Databank at Bar Ilan, an outgrowth of the earlier Responsa Project, a collection of rabbinical responses to questions on points of Jewish law; and the Thesaurus Linguae Graecae at the University of California at Irvine. This is by no means an exhaustive list, but indicates the breadth of current activity. Among corpus projects, the Brown and Lancaster-Oslo-Bergen corpora of the 1960s, and their various analogues in other languages, are now being succeeded by a new wave of larger projects, for example the British National Corpus, which will encode 100,000,000 (one hundred million) words of written and spoken British English, and the Network of European Reference Corpora. The most prominent of what I am referring to as “opportunistic projects” may be the ACL Data Collection Initiative (DCI) and the European Corpus Initiative (ECI). It is no accident that Don Walker was involved with so many of these projects.

the long run, even if we must sacrifice some modest amount of short-term convenience or efficiency in a single application.

Paradoxically, experience seems to show that the best way to ensure that one can process the text in any way one wants is to ignore processing as far as possible, and focus on saying what one thinks the text *is*. That is, one needs to find a *declarative* way of representing the text, not a *procedural* way. This involves adding a level of indirection to processing, and so is sometimes disparaged as inefficient, but it's very important.

The basic problem of putting text into computers thus turns out to be that one must find a representation of the text which captures the essentials of the text, and omits only the aspects one agrees to believe are negligible. In the practice of the forty-five years during which practitioners have been creating machine-readable texts for research purposes, one can identify some elements of a consensus regarding what is involved in such a representation:

First, it is not enough to transcribe just the characters of the text; it is necessary to be able to include further information in the electronic text as well; this control information should ideally be readily distinguished from the text itself. Borrowing a term from traditional publishing, one can distinguish *markup* (the control information) from *content*.<sup>2</sup>

By means of explicit markup or otherwise, electronic representations of text must solve five problems:

They must find a method of representing the characters or symbols of the text. This is relatively simple in the case of the characters of the Latin alphabet, the Arabic numerals, and common punctuation marks; it is less simple for accented characters, special symbols, and scripts other than the Latin alphabet, because these are not well supported by common data processing hardware or software. The situation is improving of late, with the development of ISO 10646 and Unicode, which provide a standard and very large repertoire of scripts and characters, but even with these standards it will still be necessary to find ways to represent non-standard symbols and characters (e.g., the special symbols of a personal shorthand invented by the writer of a manuscript, or non-standard characters omitted from ISO 10646 because they are non-standard).

They must represent, or choose to ignore, the overall logical and typographic structure of the text, including things like act and scene divisions and at least some phenomena like emphasis, quotation, bibliographic citation, and annotation. The history of typography offers persuasive evidence that these phenomena are important enough to thinking about texts that generations of scribes, authors, and typesetters have been forced to find print representations for them. Electronic representations of text would ignore the history of typography at their peril.

---

<sup>2</sup>There are occasional efforts to argue that markup is not necessary, and indeed is actively harmful. Perhaps the most widely known proponent of this view at the moment is Michael Hart of Project Gutenberg, which distributes ASCII-encoded public-domain texts by means of anonymous ftp servers. Each Project Gutenberg text, however, appears to contain an extensive header, giving the text's version number, file name, and date, providing a contact address, appealing for funds, and including a lengthy legal disclaimer. This header provides meta-textual information which is not strictly part of the text being transcribed, and so by definition constitutes markup of the text. Thus, even those who resist the use of formal markup languages do recognize in practice the need for markup to provide meta-information. The drawback of providing such meta-information without a formal markup scheme is that there is no convenient method to recognize automatically the boundaries between the text and the meta-information or markup.

The two-dimensional character of text in printed books and manuscripts must be reduced to a linear form in order to be represented in conventional computer file systems. This may involve changing the order of material (e.g., transcribing notes at their point of attachment), omitting material (e.g., running titles and page numbers, which are often omitted from electronic versions of texts), and finding methods of linking material which is physically separate but logically connected (e.g., end notes).

Interpretive or analytic information is often explicitly represented, as in language corpora which tag each token with its part of speech; such interpretive information may or may not be considered part of the text strictly speaking, but it is essential to certain kinds of serious work with the text. It is sometimes urged that creators of electronic texts eschew interpretation and limit themselves to the transcription of “the text itself.” On this logic, for example, some would object to procedures like the provision of part-of-speech information in language corpora like the tagged Brown and LOB corpora, on the grounds that it represents a subjective interpretation of the objective linguistic facts constituted by the wording of the texts.

As usually formulated, this objection to interpretation is intellectually problematic in itself, since no clear boundary can be drawn between interpretation and “the text itself.” The “objective linguistic facts” about the wording of the text are themselves often the subject of hot disputes among textual critics, and even the reading of the characters in a manuscript (or in a printed book) can be controversial. That is, what constitutes objective fact for one reader may seem to another to involve illicit interpretation of the text. Those who create electronic text primarily for the use of others will of course do well to distinguish between information on which there is likely to be broad agreement and information more likely to be controversial, and to allow the user of the electronic text to disregard the controversial information in a systematic way. The notion that electronic texts can be kept devoid of analytic or interpretive information, however, is chimerical: as long as researchers use electronic texts in their work, they will find it convenient to record their interim or final results in the text, for further processing later on. Any general method of text encoding must therefore provide methods for recording such interpretations.

Finally, it is often useful to record certain auxiliary information about the text, even though it may not in any way be considered part of “the text itself”: control information identifying the author and title of the text, providing a bibliographic description of the source, identifying those responsible for the electronic version, and providing other useful information about the text, is commonly recorded in electronic texts or in accompanying documentation. A strong case for providing this information within the text itself can be made from the simple observation of how frequently electronic materials are found separated from the paper documentation which originally accompanied them. In language corpora, such ancillary control information may often include characterizations of the text as a whole: e.g., demographic descriptions of the speakers in a corpus of spoken material, or classification by subject matter and text type in corpora of written materials.

From the descriptions just given, it may be observed that in practice, the researchers who have thus far put texts into electronic form have been by and large more interested in texts per se than in the details of the pages on which the texts were written. The page is one representation of a text; the electronic transcription is another. The electronic

version can of course represent the page, but it can also represent the text, without the intermediary of the page. For purposes of research with texts, what are needed are *text languages*, not *page languages*, and not just images of pages.

In emphasizing the text over the page this way, I follow the unspoken but unambiguous practice of standard practice in most textual work:

New editions, even critical editions, very rarely preserve the pagination and lineation, let alone the type face, leading, and gathering structure of earlier editions. This is only defensible if the text is not the same as the page.

Often, students are given modern-spelling editions to read; this is defensible only if the text is not the same as the accidentals of the early printings or manuscripts.

Even though any scholar recognizes the potential importance of layout, typeface, etc., and is open to their overt or subliminal influence, still it is an unusual work of scholarship in language or literature (let alone the other disciplines which concern themselves with text) in which the argument hinges on typographic or bibliographic analysis. An obvious exception, of course, are works devoted to paleography, codicology, analytic bibliography, and the history of printing and binding; practitioners in these fields will require methods of recording the details of the physical presentation of a text in a given edition or manuscript. Like other specialized information, however, this may not be of great utility to researchers in other fields.

### 3 Obstacles to Sharing Text

Machine readable texts have been in use for research for over forty-five years; this is about as long as computers have been commercially available.<sup>3</sup> In general computer-assisted projects of text analysis have historically followed a common pattern: first, the text to be analysed is recorded in electronic form, and then the analysis itself is performed and the results published.

For at least thirty years, the observation has been made that when multiple projects work with same text, the first step need not be repeated for each project: once the machine-readable text is created, it can be used for many different analyses without further encoding work. For thirty years, that is, there have been calls for machine-readable texts to be *shared*.

These calls for resource sharing, however, have been only moderately successful.

In the first place, some people don't want to share their texts. If one has gone to all the pain and trouble of creating an electronic text, and is about to perform an analysis on it, one may well be reluctant to share it with others. They may take it, perform their own analysis of it, possibly publishing before the text's creator and receiving the attendant glory. The creators of electronic texts may however prefer to retain as much glory as possible for themselves, for use when they come up for tenure, promotion, or a raise. It may be noticed that this line of argument relies on the implicit claim that relative to the analysis, the task of creating the electronic text is large and onerous. In other words, it really would save time and trouble for the research community overall, if a way could be found to make it easier and more common to share electronic texts.

---

<sup>3</sup>I take Father Roberto Busa's *Index Thomisticus* project, which began in 1948, as marking the first use of machine-readable text for research.

A second reason for the community's failure to achieve widespread text sharing is that when researchers do use each other's texts, they discover that they don't always understand them, because the methods used to encode the texts are so often idiosyncratic. This results in part from the newness of the medium. Faced with the task of representing a text in electronic form, without established conventions for the result, scholars find themselves in an Edenic position: Like Adam and Eve, the creator of an electronic text has the privilege of giving something a name, and having the name so given be *the* name of that thing. If one decrees, for example, that an asterisk is used to mark an italic word, and that a percent sign will precede and follow each personal name, and that a commercial @ sign is used to mark each place name, then that is what those things mean. The blankness of the slate gives to the encoder a kind of euphoric power, which is understandably slightly intoxicating. The result is that over the last forty years virtually every scholar who has created an electronic text has used the opportunity to wield that power and to invent a new language for encoding the text.

Electronic texts thus are, and have always been, in the position of humankind after the Tower of Babel. The result, predictably, has been pretty much what the Yahweh of Genesis had in mind: the cooperation of the research community has been hindered and delayed by the needless misunderstandings and the pointless work of translating among different systems of signs, makework that would be unnecessary if there existed an accepted common language for use in the creation of electronic texts.

Three distinct difficulties may be identified in the attempt by one researcher to use electronic texts created by someone else.

When researcher A gets a text from researcher B, first of all, A may not understand what all the special marks in it mean. If B has invented a new language, a special system of signs, that is, for this specific text, then A may find that B's text contains signifiers which are opaque because A doesn't know their significance.

The second difficulty is that once A does understand B's signs, it may become clear that the signifieds of B's text don't tell A what she wants to know. It's good that A now understand that the at-sign means a place name, but if A is interested not in place names, but rather in the use of the dative case (which B has not marked in the text), then B's text may not be as much use to A as she may have hoped before knowing what all those special marks in the text meant.

The third difficulty is that, after swallowing her disappointment and beginning to add information to B's text, specifically by marking the occurrences of the dative case, A will all too frequently find:

- that the markup language B used has no method of marking the dative case
- that it also has no provision for graceful extension of its vocabulary, and thus
- that it does not scale up well.

## 4 Solutions

These three difficulties are not equally soluble, but they are all soluble at least in part. The TEI is an attempt to solve them, as far as possible.

The second is soluble only within very restricted bounds. Without violating the autonomy of the individual researcher, it is impossible to tell each other that we all have to be interested in the dative case, instead of in place names. Within limits, however, a tenuous consensus can be formed regarding some minimum set of textual features which everyone, or almost everyone, regards as being of at least potential interest. No one should hope for too much from this consensus; the simple political fact is that very few features seem useful to absolutely everyone. Thus, I would not recommend to anyone that they should encode a text recording only the features that the universal consensus regards as useful. Almost no one would be happy with such a text: everyone regards other features as desirable, though we can reach no agreement as to what those other features are.

The first difficulty, that of understanding what it is the encoder is saying about a text, can be solved much more satisfactorily. The TEI will provide a large, thoroughly defined lexicon of signs (*tags* is the technical term) for use in marking up texts, and the published text of the Guidelines will suffice for virtually all the signifieds which workers with electronic text now record in their texts. By using this set of documented signs, one cannot guarantee that one will find the encoding work of others useful or interesting, but it can at least become probable that secondary users of the text can understand what features the encoding of a text does and does not record.

Because such a vocabulary of tags must necessarily be rather large, almost no one will be interested in using every item in it. The first task of the encoder who uses TEI markup will therefore be to make a selection among the signs defined in the scheme, and to begin making local policy decisions as to how those signs are to be used. The TEI provides, in the TEI header, a place to record those policy decisions, so that later users of the text can know what was done when the text was created.

The third difficulty, graceful extension and scale-up to more elaborate, information-rich versions of a text, the TEI handles in three ways:

First, the TEI itself is designed to be used both for rather sparse markup, which captures only a little information, and also for richer markup. That is, the TEI markup language itself scales up and down.

Second, the predefined vocabulary of the TEI includes a number of ‘built-in extensions’, by means of which new varieties of known classes may be integrated into the markup scheme without any change to its formal definition at all. For example, many markup languages (TeX, L<sup>A</sup>T<sub>E</sub>X, Script, troff, Scribe) provide tags for marking enumerated lists, bulleted lists, and possibly one or more other styles of list. In general, however, one is limited to the varieties of list foreseen in the design of the system: one cannot add a new type of list to L<sup>A</sup>T<sub>E</sub>X without modifying L<sup>A</sup>T<sub>E</sub>X. The TEI defines one basic list element, and provides a *type* attribute to allow different varieties of list (e.g., bulleted or enumerated) to be distinguished. Since the values of the *type* attribute are not constrained, a new kind of list can be introduced simply by providing a suitable value for the *type* attribute.

Third, the definition of TEI conformance explicitly envisages the formal modification of the markup language itself, in cases where this is needed. The design and integration of such modifications does require a certain technical skill, though possibly less than is required to modify L<sup>A</sup>T<sub>E</sub>X or Scribe. But it is expected that, as with those systems, a local guru will usually be found who can help the user who needs help in changing the

formal markup language.

The TEI thus builds a finite vocabulary, but explicitly plans for its growth, both by means of formal modifications to the markup language and without such modifications, by means of built-in extensions.

That is, the TEI explicitly recognizes that *no finite vocabulary is complete*.

The effort to solve the problems of interchange outlined above, by building such a scheme, began with a planning conference, held in Poughkeepsie, New York, at Vassar College in November of 1987. Thirty-one representatives of professional societies, research centers, text archives, and corpus projects met to discuss the desirability and feasibility of creating a single common scheme for encoding machine-readable texts. There was a clear consensus that such a scheme was both possible and desirable; somewhat to the surprise of the organizers, this view was shared even by the participants responsible for several of the large existing archives of electronic text, many of which have thousands of dollars and tens of staff years invested in their own locally developed encoding schemes.

At the meeting, three organizations active in the application of computers to natural-language and textual material agreed to sponsor an effort to develop a new text encoding scheme, suitable for use both in local processing and as an interchange language between sites which preferred to use their own locally developed markup languages for local processing. These were the Association for Computers and the Humanities, which under the leadership of Nancy Ide had sponsored the planning conference; the Association for Literary and Linguistic Computing, represented by both its permanent secretary Susan Hockey and its president Antonio Zampolli; and the Association for Computational Linguistics, represented by Don Walker and his colleague at Bellcore Robert Amsler. The associations in question named the individuals mentioned, together with the author, as delegates to a Steering Committee for the TEI, which began to meet almost immediately after the Planning Conference.

The Steering Committee, in turn, named the author as editor (later Lou Burnard of Oxford University Computing Service was named as associate editor), with the responsibility of planning and coordinating the work of the project, sought and received funding from the National Endowment for the Humanities, the Andrew W. Mellon Foundation, and the Commission of the European Communities (now the European Union), and invited other professional societies to join in an Advisory Board. The Advisory Board met in February, 1989, reviewing and approving the overall planning and design work done to that time. Following a plan for division of labor enunciated at the planning meeting, four working committees were appointed, with the task of address problems of:

- text documentation (especially bibliographic control information and the like)
- text representation
- text analysis and interpretation
- metalinguistic issues and syntax of the encoding scheme

Of these, the first committee had the most clearly circumscribed area of responsibility, and the second and third had an essentially unbounded scope of activity. The slightly

artificial distinction between representation and interpretation of a text was drawn for reasons of practical convenience; as a rule of thumb, the text representation committee was to be responsible for developing markup capable of recording the textual features signaled overtly (e.g., by italics, boldface, or special layout) by conventional printed books, while the committee on analysis and interpretation dealt with everything else which might be thought useful. The latter committee was instructed to concentrate its initial work on the problems of linguistic analysis, both because linguistic analysis seemed more successfully formalized than other textual disciplines and because linguistic understanding is a precondition of so many other areas of textual work.

The working committees met in 1989 and 1990 and the result of their labors was released in June of 1990 as TEI document TEI P1 ("public proposal no. 1"). In 300 letter-sized pages, this draft covered issues of characters and character-set documentation, defined a header for in-file bibliographic description of electronic texts and documentation of the encoding practices used in them, described SGML markup for a large set of features common to many text types and for the provision of analytic and interpretive information with particular reference to linguistic analysis, sketched SGML tag sets for corpora, literary texts, and dictionaries, and defined methods of extending the TEI tag sets.

After the publication of TEI P1, work immediately began on its extension and revision, and work groups were appointed to work on specialized topics such as:

- character sets
- textual criticism
- hypertext and hypermedia
- formulae, tables, figures, and graphics
- language corpora
- manuscripts and codicology
- verse
- drama and other performance texts
- literary prose
- linguistic description
- spoken text
- literary studies
- historical studies
- printed dictionaries
- machine lexica

- terminological data

These work groups met over a period of two years, and the resulting draft, TEI P2, was issued chapter by chapter beginning in early 1992 and continuing through the end of 1993. At that time all the published chapters were revised, several essential new chapters were added, and the resulting cumulative document was published in the first half of 1994 under the document number TEI P3. This version of the Guidelines has grown from 300 pages to 1300 pages, in part by the addition of an alphabetical reference list of SGML tags and in part by the addition of a great deal of new material. The table of contents for TEI P3 is this:

- Part I: Introduction

- 1 About These Guidelines
- 2 A Gentle Introduction to SGML
- 3 Structure of the TEI Document Type Definition

- Part II: Core Tags and General Rules

- 4 Characters and Character Sets
- 5 The TEI Header
- 6 Elements Available in All TEI Documents
- 7 Default Text Structure

- Part III: Base Tag Sets

- 8 Prose
- 9 Verse
- 10 Drama
- 11 Transcriptions of Speech
- 12 Print Dictionaries
- 13 Terminological Databases

- Part IV: Additional Tag Sets

- 14 Linking, Segmentation, and Alignment
- 15 Simple Analytic Mechanisms
- 16 Feature Structures
- 17 Certainty and Responsibility
- 18 Transcription of Primary Sources
- 19 Critical Apparatus
- 20 Names and Dates
- 21 Graphs, Networks, and Trees

- 22 Tables, Formulae, and Graphics
  - 23 Language Corpora
- Part V: Auxiliary Document Types
  - 24 The Independent Header
  - 25 Writing System Declaration
  - 26 Feature System Declaration
  - 27 Tag Set Documentation
- Part VI: Technical Topics
  - 28 Conformance
  - 29 Modifying the TEI DTD
  - 30 Rules for Interchange
  - 31 Multiple Hierarchies
  - 32 Algorithm for Recognizing Canonical References
- Part VII: Alphabetical Reference List of Classes
  - 33 Element Classes
  - 34 Entities
  - 35 Elements
- Entities, and Elements
- Part VIII: Reference Material
  - 36 Obtaining the TEI DTD
  - 37 Obtaining TEI WSDs
  - 38 Sample Tag Set Documentation
  - 39 Formal Grammar for the TEI Interchange-Format Subset of SGML

The design goals for the project were early formulated thus: the TEI encoding scheme should be:

1. sufficient for the needs of research
2. simple, clear, and concrete
3. usable without special software
4. rigorous and efficient to process
5. extensible
6. conformant to existing and emerging standards

These goals have not all been met in equal measure: the very size and subtlety required of the scheme by the first goal is partly at odds with the demand of the second goal that the scheme be simple, for example. In some measure, however, all of these goals have found some reflection in the final specification of the TEI encoding scheme.

- The list of topics given above, and the broad base of researchers who participated in the development of the Guidelines provide the best indication of the effort to ensure that the TEI Guidelines would suffice to meet the needs of most researchers.
- In the interests of concreteness, the TEI formulated not general advice on the construction of SGML tag sets, but a concrete TEI DTD which can be used as is for the vast majority of research projects using electronic text.
- Because SGML is human-readable, software-independent, and requires no non-ASCII characters, TEI-encoded texts can in principle be used without special-purpose software, and interested projects can develop their own software to process TEI-encoded texts; experience has shown, however, that work with TEI texts is materially aided by the use of SGML-aware software. This is particularly true of texts with complex encoding. To that extent, the third goal might plausibly be regarded as having been achieved only in part.
- Since the TEI scheme is formulated using SGML, it provides an explicit and rigorous document grammar and defines a tree-structured model of text (extended with pointers to allow the representation of directed graphs) which lends itself to efficient manipulation. To simplify the task of ad hoc software development, the TEI defines an “interchange format” which restricts the syntax of SGML to a manageable subset of the full syntax, which is thought by some to be marred by an excessive number of special cases and ad hoc rules.
- Extension of the TEI tag set is explicitly allowed in TEI-conformant texts — although this complicates the life of software developers materially and may make interchange more difficult, and so is not actively recommended.
- The standards most relevant to text encoding are ISO 8879, which defines SGML, and the various character-set standards. SGML conformance is a condition of TEI conformance, but for pragmatic reasons no single standard character set is mandated for TEI encoded texts.

## 5 The Choice of SGML

As noted, the TEI uses SGML as the basis for its encoding scheme; this section describes the basis for that choice.

First of all, SGML is non-proprietary, an international standard formulated by ISO (the International Organization for Standardization) and thus not within the control of any one software developer. This helps ensure the vendor- and platform-independence of SGML applications and of SGML-encoded data. With SGML, there is no user lock-in to specific systems; information is owned by the user, not by the propriety systems used

to manipulate it. This is sufficiently important for industry to have led to wide adoption of SGML for strategic data; it is even more important for the research community, since computer systems commonly have lives measured in years, while major literary and linguistic research projects have lives measured in decades. Even for projects of shorter duration than the Oxford English Dictionary or its various counterparts in other languages, longevity is a major issue. Work in the textual disciplines may remain relevant and important for decades or centuries; when that work takes the form of electronic texts or work with such texts, it is important that the electronic forms of the texts remain usable for a much longer life span than any software has ever yet possessed.

Second, SGML provides a reasonably good model of text: fundamentally, it allows text to be represented in a labeled tree structure, with extensions to allow pointing and the creation of directed or undirected graph structures. A variety of mechanisms are available for handling information which does not fit well into a purely hierarchical model (discussed at length in one chapter of the Guidelines). SGML is general, in contrast to markup languages like TeX or troff, which are focused on the production of printed output. It is extensible, in contrast to schemes like the Office Document Architecture (later renamed the Open Document Architecture) which do not allow for user extensions to the markup language. SGML-based markup languages are generally declarative, rather than procedural, and SGML encourages the use of analytic or descriptive, rather than appearance-oriented or presentational, markup. This helps achieve the reusability of SGML data.

## 6 The Use of SGML

The TEI encoding scheme is defined as an *application* of SGML, and its formal specification takes the form of an SGML “document type declaration.” This specification is characterized by:

- an emphasis on logical, rather than physical, structure of the text, on texts rather than on pages, for the reasons described above
- the frequent application of Occam’s Razor; for example, in the provision of a single tag for lists, with an attribute to specify the type, rather than separate tags for ordered, bulleted, and simple lists
- a modular architecture which groups tags into easily understood sets, which may be combined more or less freely for use with particular texts
- the explicit provision of methods of adding new tags, and even new tag sets, to the encoding scheme, so as to ensure that the TEI markup language remains open to improvement and extension

Particular attention has also been paid to ensuring that information of varying types can be included in the same document, and that documents can be gradually enriched by the addition of new information and analysis, without the new information getting in the way of the old. SGML software can readily ignore the markup not of interest to the user at any given moment, effectively filtering the document into a form suitable for the

particular task in hand. It is possible using the TEI scheme, for example, to combine in a single document

- orthographic transcription of the text
- pointers to a digital or analog recording of a speech signal or a videotape of an event
- markup of proper nouns, dates, times, etc.
- part-of-speech tagging
- analysis of surface syntactic structure, including multiple analyses of ambiguous structures
- analysis of the discourse structure
- cross references to other material on the same topic
- links to figures and graphics stored in any suitable notation (which need not be SGML)

A simple example may be used to show what the TEI scheme looks like in practice; most SGML-aware display software, however, will not show the tags to the user in this form, instead using font, type size, and layout guided by user-defined style sheets to signal the nature of the information being displayed.

A TEI-encoded version of Franklin Delano Roosevelt's first inaugural address, for example, might look like this:

```
<!DOCTYPE TEI.2 system 'tei2.dtd' [
<!ENTITY % TEI.prose 'INCLUDE' >
<!ENTITY wsd.en SYSTEM 'teien.wsd' SUBDOC>
]>
<TEI.2>
<teiHeader>
<fileDesc>
<titleStmt>
    <title>First Inaugural Address: An Electronic Version.</>
    <author>Franklin Delano Roosevelt.</>
    <respStmt><resp>tagged from the Project Gutenberg
                edition by</>
        <name>C. M. Sperberg-McQueen</></>
<publicationStmt>
    <authority>C. M. Sperberg-McQueen</authority>
    <pubPlace>Chicago</>
    <availability><p>This electronic text may be freely
                redistributed; it should not however be confused with
                the Project Gutenberg version of the same text, from
                which this version derives in part. The inaugural
                speech itself is in the public domain.</availability>
<date>1994</>
```

```
</publicationStmt>
<sourceDesc>
<bibl><title>"The only thing we have to fear. . . is fear
itself." President Franklin Delano Roosevelt's First
Inaugural Speech</title> [Originally delivered March 4th,
1933] ([Champaign, IL]: Project Gutenberg, 1994) [file
fdr10.txt]</bibl>
</sourceDesc>
</fileDesc>

<encodingDesc>
<projectDesc>
    <p>This tagged version of Roosevelt's inaugural was prepared
    as a demonstration of SGML tagging by C. M. Sperberg-McQueen.
    The untagged text from which it derives was produced by
    Project Gutenberg.
<editorialDecl>
    <correction status=unknown method=silent><p>Corrected by
        CMSMcQ against the text of the speech as given in Henry
        Steele Commager's <title>Documents of American History.
        </title>
    </p></correction>
</editorialDecl>
</encodingDesc>

<profileDesc>
<langUsage>
    <language id=en>U.S. English</language>
</langUsage>
</profileDesc>

<revisionDesc>
<list>
<item>26 March 1994 : CMSMcQ : complete header, tag
paragraphs of text, reformat paragraphs.
<item>10 March 1994 : CMSMcQ : add skeleton file in TEI form,
begin tagging header.
</list>
</revisionDesc>
</teiHeader>

<text>
<front>
<titlePage>
<docTitle>
<titlePart>Inaugural Speech of Franklin Delano Roosevelt</>
<titlePart>Given in Washington, D.C.</>
</docTitle>
<docDate>March 4th, 1933</docDate>
</front>
<body>
```

```
<p>President Hoover, Mr. Chief Justice, my friends:
```

```
<p>This is a day of national consecration, and I am certain  
that my fellow-Americans expect that on my induction into  
the Presidency I will address them with a candor and a  
decision which the present situation of our nation impels.
```

```
<p>This is pre-eminently the time to speak the truth, the  
whole truth, frankly and boldly. Nor need we shrink from  
honestly facing conditions in our country today. This  
great nation will endure as it has endured, will revive and  
will prosper.
```

```
<p>So first of all let me assert my firm belief that the  
only thing we have to fear is fear itself — nameless,  
unreasoning, unjustified terror which paralyzes needed  
efforts to convert retreat into advance. ...
```

```
<p>In this dedication of a nation we humbly ask the  
blessing of God. May He protect each and every one of us!  
May He guide me in the days to come!
```

```
</body>  
</text>  
</TEI.2>
```

The document begins with an SGML document type declaration, indicating that the main DTD is found in a system file called “tei2.dtd”; on the second and third lines, entity declarations identify the identifiers “TEI.prose” and “wsd.en” with, respectively, the string “INCLUDE” and the system file “teien.wsd”. The former indicates that the TEI base tag set for prose is to be included; the latter identifies an externally stored writing system declaration, which in this case documents the language (English) and character set used to encode the text. The string “]>” on the fourth line of the example ends the document type declaration.

The document instance itself begins on the fifth line. Each SGML element is delimited by a start-tag and an end-tag, themselves delimited by angle brackets or angle-bracket-slash and angle bracket. The “<tei.2>” on line 5 and the “</tei.2>” on the last line of the example show the beginning and end of the entire document instance. The root element, <TEI.2>, contains in turn two subelements: a TEI header, tagged <teiHeader>, and a <text>. The text itself contains merely a series of paragraphs, tagged <p>; the TEI header, on the other hand, has a fairly elaborate substructure used to document the electronic text, including its bibliographic source and the encoding practices used in creating it.

The allowable content (i.e., the syntax) and the semantics of the elements like <TEI.2>, <teiHeader>, and <p> are given by the TEI Guidelines, as part of the predefined vocabulary of SGML elements provided by the TEI encoding scheme.

The TEI defines a single unified encoding scheme which is scalable, allowing both very light text markup and extremely dense, information-rich markup. It provides explicit support for analysis of the text, without requiring adherence to any particular

linguistic approach or other theory,' and allowing the peaceful coexistence of many different types of analysis. Using standard SGML techniques, it makes possible the linkage of text to speech or other non-textual data at any desired level of granularity. With its wealth of flexible analytic mechanisms and its support for information filtering, the TEI encoding scheme provides a computationally tractable representation of rich text which has few serious competitors within or outside the SGML community. Above all, the work of the many volunteers on its work groups has ensured that the TEI defines a compendious inventory of textual phenomena of interest to researchers, for the description of the physical, formal, rhetorical, linguistic, and other aspects of the text.

## 7 Conclusion

By providing a common public vocabulary for text markup, we will have taken one major step toward making electronic texts as important and useful as they ought to be, but only one step. Other steps are still required.

First of all, we must as a community make a serious commitment to allowing reuse of our electronic texts. This will require either a massive upsurge in the incidence of altruism or much stronger conventions for the citation of electronic texts, and giving credit for the creation of electronic materials, both in bibliographic practice and at promotion, tenure, and salary time.

Second, we must cultivate a strict distinction between the format of our data and the software with which we manipulate it, because software is short-lived, but our texts are, or should be, long-lived. Our paper archives are full of documents fifteen or twenty years old, or 150 to 200 years old, or even 1500 or 2000 years old. But I cannot think of a single piece of software I can run which was written even 10 years ago. To allow our texts to survive, we must separate them firmly from the evanescent software we use to work on them. SGML and other standards encourage such a distinction, but proprietary products typically obscure it: in some operating systems, every document is tied, at the operating system level, to a single application — precisely the wrong approach, from this point of view.

Third, we need to cultivate better, more intelligent software, with better understanding of the nature of text structures, in order to make the texts contained in our archives more useful in our work.

Finally, we need if possible to come to a richer consensus about the ways in which we encode texts: we should try to move beyond an agreement on syntax and achieve more unity on the specific features of text which are widely useful. Such a consensus will make the TEI less of merely syntactic convention and more of a real common language.

The TEI's contribution to the success of electronic textual research will, I hope, be that it provides us with a common language, to allow us to escape our post-Babel confusion. As the list just concluded makes clear, such a common language is not all we need. But as the Yahweh of Genesis says: "If as one people speaking the same language they have begun to do this, then nothing they plan to do will be impossible for them." [Gen. 11:6, NIV]

## References

- [1] Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing (ALLC), *Guidelines for Electronic Text Encoding and Interchange*, TEI P3, ed. C. M. Sperberg-McQueen and Lou Burnard, Chicago: Oxford, Text Encoding Initiative, 1994.

# Discrimination Decisions for 100,000-Dimensional Spaces

William A. Gale  
Kenneth W. Church  
David Yarowsky

AT&T Bell Laboratories

*e-mail:* [kwc@research.att.com](mailto:kwc@research.att.com)

## Abstract

Discrimination decisions arise in many natural language processing tasks. Three classical tasks are discriminating texts by their authors (author identification), discriminating documents by their relevance to some query (information retrieval), and discriminating multi-meaning words by their meanings (sense discrimination). Many other discrimination tasks arise regularly, such as determining whether a particular proper noun represents a person or a place, or whether a given word from some teletype text would be capitalized if both cases had been used.

We (1993) introduced a method designed for the sense discrimination problem. Here we show that this same method is useful in each of the five text discrimination problems mentioned.

We also discuss areas for research based on observed shortcomings of the method. In particular, an example in the author identification task shows the need for a robust version of the method. Also, the method makes an assumption of independence which is demonstrably false, yet there has been no careful study of the results of this assumption.

## 1 Introduction

Statistical methods are being applied to more and more problems in natural language. Although there has been a long tradition of statistical work in natural language (e.g., Zipf 1932, Yule 1944, Mosteller and Wallace 1964, Salton and Yang 1973, Harris 1968), there has recently been a revival of interest in statistical approaches to natural language because computer readable text is becoming easier to obtain in large quantities. Just ten years ago, the one million word Brown Corpus (Francis and Kučera, 1982) was considered large. These days, a corpus has to be at least ten times larger in order to be considered large. And some researchers are using corpora that are a hundred times larger than the Brown Corpus.

There are a number of well-known applications of discrimination techniques in natural language processing, especially information retrieval and author identification. A

number of other problems can be addressed with very similar discrimination techniques, especially the very important problem of sense disambiguation. We have also applied similar discrimination techniques to restore capitalization in upper-case only text and to distinguish names of places from names of people.

It is an interesting question whether techniques that were developed several decades ago will continue to scale up as we continue to look at larger and larger problems. One might wonder if it is appropriate to expect a discrimination technique that was developed on a “small” problem such as the Federalist Papers to work on a “large” corpus of some tens or hundreds of millions of words of text. Most of these discrimination techniques use a  $V$ -dimensional space, where  $V$  is size of the vocabulary. The vocabulary and the dimensionality of the problem grow rapidly as we consider larger and larger corpora. The old Federalist collection contains a mere  $N = 208,304$  tokens and a vocabulary of only  $V = 8663$  types; more modern corpora contain some tens of millions of tokens and hundreds of thousands of types.

Therefore we must find discrimination techniques for dealing with spaces having about 100,000 dimensions. These methods can either be “direct,” not reducing the number of dimensions of the space, or “indirect,” reducing the number of dimensions to some manageable number. We find that Bayesian decision analysis can be used in a direct fashion for each of the problems we examine. Indirect approaches have been necessary due to computing constraints until recently, so there is some heuristic experience with them. However, principled study of indirect approaches has been possible only since direct approaches could be implemented, which is to say recently, so little is known about them.

## 2 Discrimination Problems in Natural Language Processing

Text discrimination problems begin by specifying a corpus, a collection of documents such as the Federalist Papers, newswire stories collected from Associated Press (AP) over a few years, the official record of the Canadian parliamentary debates, or a set of encyclopedia articles. Documents are represented as a sequence of tokens, e.g., words, punctuation, type-setting marks, and delimiters to mark sentences and paragraphs.

In the training phase, we are given two (or more) sets of documents and are asked to construct a discriminator which can distinguish between the two (or more) classes of documents. These discriminators are then applied to new documents during the testing phase. In the author identification task, for example, the training set consists of several documents written by each of the two (or more) authors. The resulting discriminator is then tested on documents whose authorship is disputed. In the information retrieval application, a query is given and the training set consists of a set of one or more documents relevant to the query and a set of zero or more irrelevant documents. The resulting discriminator is then applied to all documents in the library in order to separate the more relevant ones from the less relevant ones. In the sense disambiguation case, the 100-word context surrounding instances of a polysemous word (e.g., *bank*) can be treated very much like a document, as we will see.

There is an embarrassing wealth of information in the collection of documents that could be used as the basis for discrimination. To date, most researchers using statistical techniques have tended to treat documents as “merely” a bag of words, and

have generally tended to ignore much of the linguistic structure, especially dependencies on word order and correlations between pairs of words. The collection of documents can then be represented as a term by document matrix, where each cell counts the number of times that a particular term appears in a particular document. Since there are  $V \approx 100,000$  terms, the term by document matrix contains a huge amount of information, even allowing for the fact that the matrix is quite sparse and many of the cells are empty.

One approach to these problems has been Bayesian discrimination analysis. Mosteller and Wallace (1964, section 3.1) used the following formula to combine new evidence (e.g., the term by document matrix) with prior evidence (e.g., the historical record) in their classic authorship study of the Federalist Papers.

$$\text{final odds} = (\text{initial odds}) \times (\text{likelihood ratio})$$

For two groups of documents, the equation becomes

$$\frac{P(\text{class}_1)}{P(\text{class}_2)} = \frac{p(\text{class}_1)}{p(\text{class}_2)} \times \frac{L(\text{class}_1)}{L(\text{class}_2)}$$

where  $P$  represents a final probability,  $p$  represents an initial probability, and  $L$  represents a likelihood. Similar equations can be found in textbooks on information retrieval (e.g., Salton 1989, equation 10.17).

The initial odds depend on the problem. In the author identification problem, for example, the initial odds are used to model what we know about the documents from the various conflicting historical records. In the information retrieval application, the user may have a guess about the fraction of the library that he or she would expect to be relevant; such a guess could be used as the prior. The objectivity of the prior depends on the problem. In the author identification case, it is largely subjective. For the information retrieval problem, a baseline probability could be established from past experience in the number of relevant documents found. For many other problems, including spelling correction, sense disambiguation, and other problems discussed here, the prior can be quite objective and very useful.

It is common practice to decompose the likelihoods into a product of likelihoods over tokens in the document (under appropriate independence assumptions):

$$\frac{L(\text{class}_1)}{L(\text{class}_2)} \approx \prod_{\text{token in doc}} \frac{Pr(\text{token}|\text{class}_1)}{Pr(\text{token}|\text{class}_2)}$$

The crucial ingredients for this calculation are the probabilities of each term in the vocabulary *conditional* on the document being from a given class. These conditional probabilities have been computed in a number of different ways depending on the application and the study. In the next section we will introduce a novel method of calculating these conditional probabilities. The method was originally designed for the sense disambiguation application, though we have found that the method can be used “off-the-shelf” to produce results that are comparable to (though perhaps not quite as good as) methods that have been highly tuned over many years for a particular problem.

### 3 Sense Discrimination: An Example of the Approach

Consider, for example, the word *duty*, which has at least two quite distinct senses: (1) a tax and (2) an obligation. Three examples of each sense are given below in Table 1.

Table 1: Sample Concordances of *duty* (split into two senses)

Sense	Examples (from Canadian Hansards)
tax	companies paying <b>duty</b> and then claiming a refund a countervailing <b>duty</b> of 29.1 per cent on canadian states imposed a <b>duty</b> on canadian saltfish last year
obligation	is my honour and <b>duty</b> to present a petition duly beyond the call of <b>duty</b> ? SENT i know what time addition , it is my <b>duty</b> to present the government 's

Sense disambiguation has been recognized as a major problem in natural language processing research for over forty years. There has been a considerable body of work on the subject, but much of the work has been stymied by difficulties in acquiring appropriate lexical resources (e.g., semantic networks, annotated corpora, dictionaries, thesauruses, etc.). In particular, much of the work on qualitative methods has had to focus on “toy” domains since currently available semantic networks generally lack the broad coverage that would be required to address a realistic problem. Similarly, much of the work on quantitative methods has had to depend on small amounts of hand-labeled text for testing and training.

We achieved considerable progress as reported in (Gale et al., 1993) by taking advantage of a new source of testing and training materials for studying sense disambiguation methods. Rather than depend on small amounts of hand-labeled text, we used relatively large amounts of parallel text, text such as the Canadian Hansards, which are available in multiple languages. The translation can often be used in lieu of hand-labeling. For example, the two senses of *duty* mentioned above are usually translated with different French words in the French version. The tax sense of *duty* is typically translated as *droit* whereas the obligation sense is typically translated as *devoir*. Thus, we can collect a number of tax sense instances of *duty* by extracting instances of *duty* that are translated with *droit*, and we can collect a number of obligation instances by extracting instances that are translated with *devoir*. In this way, we were able to acquire considerable amounts of testing and training material for study of quantitative methods. More details on the preparation of the testing and training materials can be found in (Gale and Church, 1991a, 1991b).

The availability of this testing and training material has enabled us to develop quantitative disambiguation methods that achieve 92 percent accuracy in discriminating between two very distinct senses of a noun such as *duty*. While the 8% error rate appears to be about half that of disambiguation methods published before we began the work, it is even more important that the proposed method has a better potential of scaling up to handle realistic size vocabularies of tens of thousands of ambiguous words. There have been several other studies of sense disambiguation recently (Brown et al., 1991), (Dagan, Itai, and Schwart, 1991), (Hearst, 1991), (Zernik, 1992), (Yarowsky, 1992), and (Leacock et al., 1993).

We made a number of studies, most of which focus on the six English nouns shown in Table 2 (below). This table also shows the two French translations and an English gloss of the relevant sense distinction.

Table 2: Six Polysemous Words

English	French	sense	N	% correct
duty	droit	tax	1114	97
	devoir	obligation	691	84
drug	médicament	medical	2992	84
	drogue	illicit	855	97
land	terre	property	1022	86
	pays	country	386	89
language	langue	medium	3710	90
	langage	style	170	91
position	position	place	5177	82
	poste	job	577	86
sentence	peine	judicial	296	97
	phrase	grammatical	148	100
Average				90
Average	(with prior)			92

For two senses, the Bayesian equation mentioned above becomes:

$$\frac{P(sense_1)}{P(sense_2)} = \frac{p(sense_1)}{p(sense_2)} \times \frac{L(sense_1)}{L(sense_2)}$$

where  $p$ ,  $P$  and  $L$  are the initial probability, the final probability and likelihood, as before. The initial probabilities are determined from the overall frequencies of the two senses in the corpus. As in other large dimension discrimination problems, the likelihoods are decomposed into a product over tokens:

$$\frac{L(sense_1)}{L(sense_2)} \prod_{\text{token in context}} \frac{Pr(\text{token}|sense_1)}{Pr(\text{token}|sense_2)}$$

As mentioned above, this model ignores a number of important linguistic factors such as word order and collocations (correlations among words in the context). Nevertheless, there are  $2V \approx 200,000$  parameters in the model. It is a non-trivial task to estimate such a large number of parameters, especially given the sparseness of the training data. The training material typically consists of approximately 12,000 words of text (100 words of context for 60 instances of each of two senses). Thus, there are more than 15 parameters to be estimated for each data point. Clearly, we need to be fairly careful given that we have so many parameters and so little evidence.

The conditional probabilities,  $Pr(\text{token}|sense)$ , can be estimated in principle by selecting those parts of the entire corpus which satisfy the required conditions (e.g., 100-word contexts surrounding instances of one sense of *duty*), counting the frequency of

each word, and dividing the counts by the total number of words satisfying the conditions. However, this estimate, which is known as the maximum likelihood estimate (MLE), has a number of well-known problems. In particular, it will assign zero probability to words that do not happen to appear in the sample. Zero is not only a biased estimate of their true probability, but it is also unusable for the sense disambiguation task (and for quite a number of other applications). In addition, MLE also produces poor estimates for words that appear only once or twice in the sample. In another application (spelling correction), we have found that poor estimates of context are worse than none; that is, at least in this application, we found that it would be better to ignore the context than to model it badly with something like MLE (Gale and Church, 1990).

The method derived in the next section was introduced by Gale, Church and Yarowsky (1993) and uses the information from the entire corpus in addition to information from the conditional sample in order to avoid these problems. We will estimate  $Pr(token|sense)$  by interpolating between word probabilities computed within the 100-word context and word probabilities computed over the entire corpus. For a word that appears fairly often within the 100-word context, we will tend to believe the local estimate and will not weight the global context very much in the interpolation. Conversely, for a word that does not appear very often in the local context, we will be much less confident in the local estimate and will tend to weight the global estimate somewhat more heavily. The key observation behind the method is this: the entire corpus provides a set of well measured probabilities which are of unknown relevance to the desired conditional probabilities, while the conditional set provides poor estimates of probabilities that are certainly relevant. Using probabilities from the entire corpus thus introduces bias, while using those from the conditional set introduce random errors. We seek to determine the relevance of the larger corpus to the conditional sample in order to make this trade off between bias and random error.

The interpolation procedure makes use of a prior expectation of how much we expect the local probabilities to differ from the global probabilities. Mosteller and Wallace "expect[ed] both authors to have nearly identical rates for almost any word" (p. 61). In fact, just as they had anticipated, we have found that only 2% of the vocabulary in the Federalist corpus has significantly (3 standard deviation) different probabilities depending on the author. Moreover, the most important words for the purposes of author identification appear to be high frequency function words. Our calculations show that *upon*, *of* and *to* are strong indicators for Hamilton and that *the*, *and*, *government* and *on* are strong indicators for Madison. These are all high frequency function words (at least in these texts), with the exception of *government*, which is, nonetheless, extremely common and nearly devoid of content.

In contrast, we expect fairly large differences in the sense disambiguation application. For example, we find that the tax sense of *duty* tends to appear near one set of content words (e.g., *trade* and *lumber*) and that the obligation sense of *duty* tends to appear near quite a different set of content words (e.g., *honour* and *order*), at least in the Hansard corpus. Approximately 20% of the vocabulary in the Hansards has significantly different probabilities near *duty* than otherwise. In short, the prior expectation depends very much on the application. In any particular application, we set the prior by estimating the fraction of the vocabulary whose conditioned probabilities differ significantly from the global probabilities. Thus the same interpolation procedure is used for all of the

applications discussed here.

## 4 The Interpolation Procedure

Let the entire corpus be divided into a sample, satisfying some condition, of size  $n$ , and the residual corpus (the entire corpus less the conditional sample) of size  $N \gg n$ . Let  $a$  be the frequency of a given word in the conditional sample, and  $A$  its frequency in the residual corpus. Either of these frequencies may be zero, but not both. Let  $\pi$  represent the probability of the word given the condition establishing the sample. Before knowing the frequency of the word in either the sample or the residual corpus, we could express our ignorance of the value of  $\pi$  by the *uninformative distribution*:

$$B^{-1}\left(\frac{1}{2}, \frac{1}{2}\right)\pi^{-\frac{1}{2}}(1-\pi)^{-\frac{1}{2}}$$

where  $B(x, y)$  is the Beta function. Several variations of the method can be based on variations in the uninformative distribution. If  $A$  instances out of  $N$  independent observations relevant to the determination of  $\pi$  were found, then the distribution expressing our knowledge would become

$$B^{-1}\left(A + \frac{1}{2}, N - A + \frac{1}{2}\right)\pi^{A-\frac{1}{2}}(1-\pi)^{N-A-\frac{1}{2}}$$

While we have  $A$  out of  $N$  observations of the word in question in the residual corpus, we do not know their relevance. Thus we set as our knowledge before observing the conditional sample the distribution:

$$\begin{aligned} p(\pi) &= rB^{-1}\left(A + \frac{1}{2}, N - A + \frac{1}{2}\right)\pi^{A-\frac{1}{2}}(1-\pi)^{N-A-\frac{1}{2}} \\ &\quad + (1-r)B^{-1}\left(\frac{1}{2}, \frac{1}{2}\right)\pi^{-\frac{1}{2}}(1-\pi)^{-\frac{1}{2}} \end{aligned}$$

where  $0 \leq r \leq 1$  is interpreted as the relevance of the residual corpus to the conditional sample. When  $r = 0$ , this gives the uninformative distribution, while if  $r = 1$ , it gives the distribution after observing the residual corpus. Another way of reading this is that with probability  $r$  we are expecting an observation in line with the residual corpus, but that with probability  $1 - r$  we won't be surprised by any value.

The joint probability of observing  $a$  out of  $n$  instances of the word in question in the conditional sample and that the conditional probability is  $\pi$  is then

$$\begin{aligned} p(\pi, a) &= \binom{n}{a} \left\{ rB^{-1}\left(A + \frac{1}{2}, N - A + \frac{1}{2}\right)\pi^{A+a-\frac{1}{2}}(1-\pi)^{N-A+n-a-\frac{1}{2}} \right. \\ &\quad \left. + (1-r)B^{-1}\left(\frac{1}{2}, \frac{1}{2}\right)\pi^{a-\frac{1}{2}}(1-\pi)^{n-a-\frac{1}{2}} \right\} \end{aligned}$$

We then form

$$p(a) = \int_0^1 p(\pi, a) d\pi$$

and

$$p(\pi|a) = \frac{p(\pi, a)}{p(a)}$$

which is then integrated to give

$$E(\pi|a) = \int_0^1 \pi p(\pi|a) d\pi = \frac{r \frac{B(A+a+\frac{1}{2}, N-A+n-a+\frac{1}{2})}{B(A+\frac{1}{2}, N-A+\frac{1}{2})} + (1-r) \frac{B(a+\frac{1}{2}, n-a+\frac{1}{2})}{B(\frac{1}{2}, \frac{1}{2})}}{r \frac{B(A+a+\frac{1}{2}, N-A+n-a+\frac{1}{2})}{B(A+\frac{1}{2}, N-A+\frac{1}{2})} + (1-r) \frac{B(a+\frac{1}{2}, n-a+\frac{1}{2})}{B(\frac{1}{2}, \frac{1}{2})}}$$

This can be approximated in various ways, but it is practical to calculate it directly using the relationship

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

The parameter  $r$ , which denotes the relevance of the residual corpus to the conditional sample, can be estimated in various ways. Its basic interpretation is the fraction of words that have conditional probabilities close to their global probabilities (as estimated from the residual sample). Thus given a set of estimates of conditional probabilities, one can estimate  $r$  as the fraction of them which lie within a few standard deviations of the corresponding global probabilities. This estimate is performed using the words which are observed in the conditional sample. Alternatively  $r$  can be regarded as a free parameter of the method and adjusted to produce optimal results on a specific task. Although it could be varied for each word, we have used  $r = 0.8$  for all words in the sense disambiguation application, and  $r = 0.98$  for all words in the author identification application, based on empirical findings from the data.

## 5 Example of the Interpolation Procedure

Table 3 gives a sense of what the interpolation procedure does for some of the words that play an important role in disambiguating between the two senses of *duty* in the Canadian Hansards. Recall that the interpolation procedure requires a conditional sample. In this example, the conditional samples are obtained by extracting a 100-word window surrounding each of 60 training examples. The training examples were selected by randomly sampling instances of *duty* in the Hansards until 60 instances were found that were translated as *droit* and 60 instances were found that were translated as *devoir*. The first set of 60 are used to construct the model for the tax sense of *duty* and the second set of 60 are used to construct the model for the obligation sense of *duty*.

The column labeled “freq” shows the number of times that each word appeared in the conditional sample. For example, the count of 50 for the word *countervailing* indicates

Table 3: Selected Portions of Two Models

tax sense of <i>duty</i>			
weight*freq	weight	freq	word
285	5.7	50	countervailing
111.8	4.3	26	duties
99.9	2.7	37	u.s
73.1	1.7	43	trade
70.2	1.8	39	states
69.3	3.3	21	duty
68.4	3.6	19	softwood
68.4	1.9	36	united
58.8	8.4	7	rescinds
54	3.0	18	lumber
50.4	4.2	12	shingles
50.4	4.2	12	shakes
46.2	2.1	22	against
41.8	1.1	38	canadian
obligation sense of <i>duty</i>			
weight*freq	weight	freq	word
64	3.2	20	petitions
59.28	0.3	228	to
51	3.0	17	petition
47.6	2.8	17	pursuant
46.28	0.5	89	mr
37.8	2.7	14	honour
37.8	1.4	27	order
36	2.0	18	present
33.6	2.8	12	proceedings
31.5	3.5	9	prescription
31.32	0.9	36	house
29.7	3.3	9	reject
29.4	4.2	7	boundaries
28.7	4.1	7	electoral

that *countervailing* appeared 50 times within the 100-word window of an instance of *duty* that was translated as *droit*. This is a remarkable fact, given that *countervailing* is a fairly unusual word.

The second column (labeled “weight”) models the fact that 50 instances of *countervailing* are more surprising than 228 instances of *to* (even though 228 instances of *to* is somewhat surprising). The weights for a word are its log likelihood in the conditional sample compared with its log likelihood in the global corpus. The first column, the product of these log likelihoods and the frequencies, is a measure of the importance, in the training set, of the word for determining which sense the training examples belong to. Note that words with large scores do seem to intuitively distinguish the two senses, at least in the Canadian Hansards. The set of words listed in Table 3 under the obligation sense of *duty* is reasonable given that the Hansards contain a fair amount of boilerplate of the form: “Mr. speaker, pursuant to standing order..., I have the honour and duty to present petitions duly signed by... of my electors...”.

By interpolating between the local and global probabilities in this way, we are able to estimate considerably more parameters than there are data points (words) in the training corpus. The interpolation procedure assumes that one selection of natural language is roughly similar to another. In this way, it becomes feasible to estimate the  $2V \approx 200,000$  parameters, one for each word in the vocabulary and each of the two senses. This is the basis for a direct approach to the problem of a high dimensional space using Bayesian decision analysis. We are, of course, assuming that conditional on being in the training sample, correlations between words are zero. Although this is a common assumption in information retrieval and author identification applications, it might be a cause of some concern. Fortunately, there are some reasons to believe that the correlations do not have a very large effect, which we will review in Section 11. But first, let us describe some other applications of discrimination methods in language. We will start with the two very well established applications, author identification and information retrieval, and then we will move on to describe some applications that we have been working on in our lab.

## 6 Author Identification

The approach we have taken to discrimination in a high dimensional space was inspired, as we have said, by the classic work in author discrimination by Mosteller and Wallace (1964). After completing the study described in the previous section, we decided to spend a little time (approximately one day) investigating how well the same methods would work on Mosteller and Wallace’s application. Although it may not be fair to compare a single day of work with a decade of research, we are excited by the fact that we could use the basic techniques “off-the-shelf” to produce nearly as good results, without spending any time tuning the methods to the particular application.

Mosteller and Wallace studied the Federalist papers, a collection of 85 documents debating the adoption of a new constitution by the American states. Of these, 51 are known to be by Hamilton (H), 14 are known to be by Madison (M), 5 by Jay (J), and 3 jointly by Madison and Hamilton (MH). The remaining 12 are of unknown authorship, but are presumed to be by either Hamilton or Madison (M?). Mosteller and Wallace

found that all twelve would be ascribed to Madison if one's prior was indifferent between Hamilton and Madison before considering this evidence.

In the previous section, we set  $r$ , the fraction of words whose conditional probabilities are similar to the global probabilities, to 0.8, based on the observation that about 20% of the vocabulary has significantly different probabilities in the conditional sample than they would otherwise have. As mentioned above, we believe that  $r$  depends fairly strongly on the application, and after a quick investigation, we decided to set  $r$  to 0.98, based on the observation that only 2% of the Federalist vocabulary has significantly different probabilities depending on the author. Thus for this problem,  $r$  is determined from the data and is not subjective.

We then built a Hamilton model from all the known Hamilton papers, and a Madison model from all the known Madison papers. These models were applied to the remaining papers, namely J + MH + M?. A positive score would indicate authorship by Madison. All of the disputed papers ( $au = M?$ ) scored positively. Thus, we reach the same conclusion as Mosteller and Wallace did. That is, assuming equal priors for the two papers, then we conclude that all of the disputed papers were more likely to have been written by Madison than by Hamilton after taking the word frequency evidence into account.

We also wanted to test the H and M papers. For each H paper, we built a model from all the Hamilton papers excepting the one in question, comparing the results to the overall Madison model. Likewise for each Madison paper we built such a cross validating model and compared results to the overall Hamilton model. Thereby, the scores for the known papers can be used to cross-check the validity of the method. Except for one mistake (document 47), all of the H papers scored negatively and all of the M papers scored positively, as they should.

Our cross-check is probably somewhat more thorough than the one that Mosteller and Wallace were able to perform forty years ago, since we now have the computer power to use a jackknifing technique and exclude each of the H and M papers from the models while they are being scored. If we had not taken this extra precaution, we would not have noticed the problem with document 47.

The one clear mistake is instructive. The problem with document 47 can be attributed to the single word *court*. The word *court* is mentioned 8 times in document 47, and in fact, document 47 is the only M document to mention the word *court* at all. And to make matters worse, *court* happens to be a favorite H word (Hamilton used it 90 times). If we remove that word, then we would find that the paper is very strongly in M's direction. Removing *court* has the effect of adding 68 to the score. Thus, instead of scoring -16 which is well within the H range, it would receive a strong M score of 42. This example suggests that we need to investigate robust approaches which would reduce the danger that one outlier could dominate the result.

## 7 Information Retrieval: Alternate Approaches

There have been quite a number of attempts to apply statistical techniques to problems in information retrieval. The *vector-space* model (Salton, 1989, section 10.1) is perhaps the most popular statistical method among researchers in information retrieval. Let  $q$  be a query, and let  $D = d_i$  be a set of documents. The information retrieval task is to sort the

documents by how similar they are to the query. According to the vector-space model, the query and the documents are represented as vectors. Each vector has  $V$  dimensions, where  $V$  is the vocabulary of terms. Thus, if the term, *strong*, for example, appears 5 times in a document, then its vector will have a magnitude of 5 along the dimension corresponding to *strong*. It is then a straightforward matter to sort vectors by a standard similarity measure such as cosine distance:

$$sim(x, y) = \frac{\sum_{i=1}^V x_i y_i}{|x| |y|}$$

where

$$|x| = \sqrt{\sum_{i=1}^V x_i^2}$$

The probabilistic retrieval model, (Salton, section 10.3), a rival to the vector-space model, sorts documents by how likely the words in the document were to have come from relevant documents  $Pr(token|rel)$ , as opposed to irrelevant documents  $Pr(token|irrel)$ . The probabilistic retrieval model shares many of the same fundamentals as the methods we have been studying, although the estimation of conditional probabilities is somewhat different in detail.

$$score(d_i) = \prod_{token \text{ in } d_i} \frac{Pr(token|rel)}{Pr(token|irrel)}$$

In our experiments, vector-space methods worked slightly better for information retrieval than our probability measure. (So naturally we tried them on the sense discrimination problem, but found them markedly inferior.) One factor that differentiates the problems is the length of query and of document. In the sense disambiguation task the query may consist of a hundred words of relevant context, while for information retrieval, queries are typically short, perhaps less than ten words.

Using either method, the greatest increase in performance for information retrieval comes from using a long query. Table 4 (below) shows some results produced by a probabilistic retrieval model. The document set consisted of AP news stories collected during April 1990. The query was the first April AP newswire story about Ryan White, a hemophiliac who died of AIDS on April 8, 1990. The table shows all stories in the collection which had a positive log likelihood score. That is, according to the model, these (and only these) stories are more likely to have been generated from the query story distribution than the global distribution.

In testing the stories, the headwords and titles were not used, so they can be examined as evidence of success in retrieval. It will be seen that the highest scoring stories are all relevant. The lower scoring stories tend to reflect part of the query, usually AIDS. No stories about Ryan White were omitted by this test. This example shows a remarkable performance for information retrieval, but it is due to using an entire story as the query, and not to the method. The vector space method does about as well on this example, although it does not have a natural cutoff, so comparison is difficult.

The information retrieval task shows that non-probabilistic methods may be better for some high dimensional problems, but that probabilistic methods can be applied “off the shelf” with competitive results.

Table 4: Probabilistic Search on “Obit-White” Story

loglike	date	headword	title
0.92	4-02	RyanWhite	AIDS Patient Ryan White Reported Near Death
0.36	4-02	RyanWhite	AIDS Victim Unconscious, Said to Be Dying from Internal ...
0.34	4-08	Bush-White	Bush Mourns Boy's Death
0.30	4-03	RyanWhite	
0.30	4-09	White-Chro	White's Struggle With AIDS Began At Age 13
0.21	4-03	RyanWhite	Ryan White, AIDS Patient, Near Death; Well-Wishers' Calls ...
0.20	4-03	RyanWhite	
0.18	4-04	RyanWhite	AIDS Patient Ryan White Remains in Critical Condition
0.18	4-09	Obit-White	Ryan White Taught Nation to Care About AIDS, Friends Say
0.18	4-04	RyanWhite	Town That Once Spurned Ryan White Joins Nation in Wishing
0.16	4-11	Bush-Health	Bush to Undergo Physical on Thursday
0.13	4-05	RyanWhite	In City That Barred AIDS Boy, an Outpouring of Sympathy
0.13	4-06	RyanWhite	Ryan White Brings Hope to Other Patients with Gifts from ...
0.13	4-10	RyanWhite	AIDS Victim Ryan White Remembered as Teacher, Student
0.12	4-07	Ryan'sLega	Unassuming Indiana Teen-ager Taught America About AIDS
0.11	4-10	RyanWhite	Hundreds Pay Respect at Funeral Home
0.11	4-11	RyanWhite	Barbara Bush, Celebrities to Attend Ryan White Funeral
0.09	4-11	RyanWhite	1,500 Say Goodbye to AIDS Victim Ryan White
0.09	4-08	White-Reax	Americans Pay Tribute to AIDS Victim Ryan White
0.09	4-12	RyanWhite	First Lady, Celebrities Attend Funeral for Young AIDS Victim
0.08	4-25	RyanWhite	Victim's Mother Lobbies for AIDS Bill
0.07	4-20	AIDSBoy	Church Bars AIDS-Infected Boy from Sunday School
0.06	4-20	DigestBriefs	
0.06	4-20	CDC-AIDS	Woman Gets AIDS Virus After Being Inseminated with Infected
0.06	4-20	AIDSBoy	Church Bars Boy With AIDS, Then Reverses Itself
0.06	4-10	Singapore	AIDS Test Required For New Maids
0.05	4-26	RyanWhite	AIDS Victim's Mother Lobbies for Spending Bill
0.05	4-22	Academic	Top 15 Finishers Listed
0.05	4-13	Students	Death of AIDS Victim Ryan White Sparks Protest
0.05	4-18	Scotus-Des	
0.04	4-04	RyanWhite	Hemophiliacs Live with Uncertainty of AIDS Infection
0.03	4-13	Kenya-AIDS	More than Four-Fifths of Kenyan Prostitutes Carry AIDS
0.03	4-16	RuralAIDS	AIDS Panel Studies Special Problems Of Rural Patients
0.03	4-05	Quotes	
0.03	4-09	Briefs	Lithuanian Declaration Should Be Withdrawn, Soviet Says

## 8 Capitalization Restoration

Although author identification and information retrieval are perhaps the two best known applications of high dimensional discrimination analysis in natural language processing, there are plenty of additional applications that are also worth mentioning. We have recently used these techniques for restoring capitalization in text that has been transmitted through a noisy channel that corrupted the case information. Some text, such as that collected from a teletype, does not preserve capitals. Also, in sentence initial position, all words are capitalized. Thus it is sometimes useful to be able to distinguish the two words by context. For many purposes, it is desirable to try to recover the capitalization information. For example, we may wish to be able to distinguish the country *Turkey* from the bird *turkey*.

We have made a few preliminary studies of this issue. It has been possible to gather large sets of training examples automatically, although considerable care was needed to avoid the multiplicity of situations in which all words are capitalized. The example of */T/t/Jurkey* is a biased example: it is our most successful model to date. When the Bayesian model of context was trained on 200 examples each of *Turkey* and *turkey* drawn from the Associated Press newswire, and tested on an additional fifty examples of each, each of the 100 test examples was correctly classified. Performance is generally closer to 90% on more typical examples.

A similar problem, in which a large dimensional model would form part of the solution is the restoration of accents. The Spanish EFE newswire deletes accents, and hardware limitations in the immediate future may create other such situations. Accent deletion can create ambiguities. If some unaccented French text contained *peche*, for instance, one would need to distinguish among *pêche*, *pèche* and *péché*.

## 9 Person or Place?

Our final example is that of distinguishing proper nouns that refer to people from proper nouns that refer to places. *Madison*, for instance, can refer to the former president of the United States or a city in Wisconsin. Since our work on this question remains preliminary, we will keep the discussion brief.

For a preliminary study, we considered discriminating city names from people's names. City names are convenient because there are a number of readily available sources of city names (e.g., almanacs, atlases). We need to be careful in selecting training material. If we simply collected all examples that we happened to have lying around, we might well end up producing a model that would be heavily biased toward the frequent items and may not reflect the generality very well. Thus we trained using sets with one example each of each name.

The models were tested on sets containing a second example of each name. We found that for this problem, as opposed to the sense discrimination problem, a narrow context of just one or two words to either side was best. This provided us with a context model.

However, for this problem, there are often strong priors about whether a name represents a person or a city. *Bush* is never a place, and in contemporary text, it is unlikely that *Madison* will refer to the former president. For each of the names we had, we made an iterative calculation to reach a probability for its representing a person. The model classifies each instance of a name assuming person or city equally likely a priori; the fraction classified as people can be taken as a prior and the group reclassified. This is an Expectation-Maximization calculation (Dempster, 1977), known to converge to a local maximum, but of unknown relation to the global maximum, or the truth. We made a series of studies by Monte Carlo techniques, and derived a calibration curve.

About three fourths of the names had just person or just city examples in training material, and errors on these groups were less than one percent on test material. Table 5 shows an example of the performance for *Dixon*, a name that is quite ambiguous, with a nearly equal distribution between person and place.

The first column in the table is the computer's assignment. The double asterisk in this column shows the one of twenty five examples in which the computer assignment

Table 5: Person/Place Assignment

Guess	Score	Text
City	-495	The people of <b>Dixon</b> are reading about
City	-893	journalist will visit <b>Dixon</b> for a week
Person	1997	lighter sentence on <b>Dixon</b> because he had
Person	1998	attorneys for <b>Dixon</b> had filed a
City **	-647	trial would prevent <b>Dixon</b> from challenging
Person	49	Bailey denied <b>Dixon</b> his constitutional
Person	5449	Kunstler said <b>Dixon</b> contends the two
City	-2776	still standing in <b>Dixon</b> that Reagan att
Person	173	the savings bank <b>Dixon</b> owned , Vernon
City	-19	miles west to <b>Dixon</b> later Wednesday
City	-4594	Reagan moved to <b>Dixon</b> with his family
City	-1057	where proud <b>Dixon</b> residents brag ,
City	-4594	Reagan moved to <b>Dixon</b> with his family
City	-4347	lived in <b>Dixon</b> from 1920 until
City	-2063	years they called <b>Dixon</b> home .
City	-2711	family lived in <b>Dixon</b> and it's the
City	-3758.	ntacted officials in <b>Dixon</b> in early 1987
City	-7736	he arrived in <b>Dixon</b> last Thursday ,
City	-2819	lived in several <b>Dixon</b> houses with his
Person	1315	official at whom <b>Dixon</b> and others in
Person	2788	estigation , which <b>Dixon</b> said could cont
Person	145	Jones credits <b>Dixon</b> with helping dev
Person	4944	Police said <b>Dixon</b> was shot in
Person	1901	a letter to <b>Dixon</b> asking if the
Person	690	sat motionless as <b>Dixon</b> read his ruling
<b>** = Incorrect</b>		

differed from that of one judge. The second column is the computer's score, negative for cities, positive for people. The final column is a small amount of context so that the reader can judge the results. By combining priors with context evidence, we reach mid-nineties for accuracy in ambiguous cases, resulting in an overall error rate of less than two percent, since the ambiguous cases are only a quarter of the total.

This use shows the need for care in gathering training material, the selection of appropriate model parameters, the use of the model to construct priors, and the utility of combining priors with the contextual information. The study needs to be extended to other kinds of places besides cities, and it needs to deal explicitly with proper nouns that are new, probably by building models from internal clues (such as morphology (e.g., *-burg*)). Nevertheless, we are quite pleased with these preliminary results.

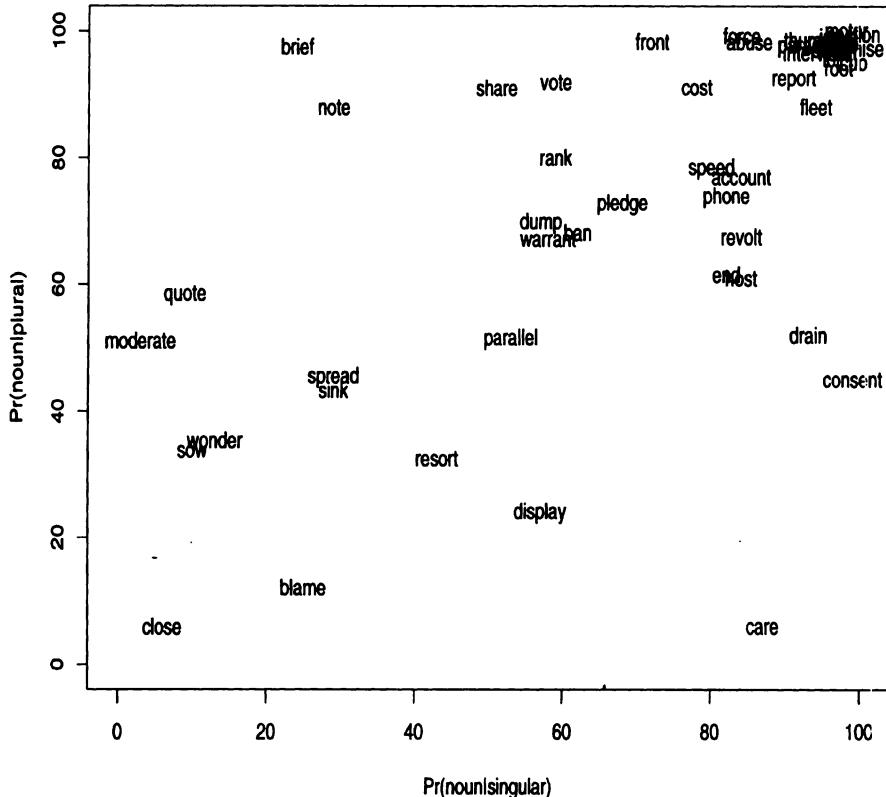
## 10 Can We Reduce the Dimensionality?

Indirect approaches to the problem of high dimensionality attempt to reduce the dimensionality.

Information retrieval work started in days when reducing dimensionality was essential to computing efficiency. Stop lists and term selection were developed as standard, if heuristic methods to reduce dimensionality. Now, however, large vocabularies can be dealt with mechanically, and the question asked of any dimension reduction scheme is how it effects precision and recall. One early method for dimension reduction was suffix removal, or stemming. Harman (1991) studied three different stemming algorithms in common use, including the simplest "S-stemmer" which just reduces English forms ending in *s* (plurals for nouns, third person singular for verbs) to their root, and two more complex schemes. She concluded "The use of the three general purpose stemming algorithms did not result in improvements in retrieval performance in the three test collections examined, as measured by classical evaluation techniques." Figure 1 may help explain why she came to this conclusion.

The quantities plotted in Figure 1 are the probabilities that a word is a noun given an apparently singular form (no *s* ending) or an apparently plural form (*s* ending). In the upper right corner, the figure shows a group of words, about 30 percent of this sample of fifty words, which are almost always nouns in either the singular or the plural. Even with these words included, the correlation is .67, low enough to show that there is considerably different usage between singular and plural forms. The remaining 70 percent of the words, however, have only a .37 correlation, showing very little relationship between usage of the singular and plural forms. Basically, the singular and plural forms usually represent dimensions which are not particularly parallel, so stemming is not much better than just dropping words arbitrarily in its dimension reduction. Conceivably, a careful study of particular nouns, locating those which were nearly always a noun in either the singular or plural forms could be used to reduce the dimensionality.

We made a brief study of the importance of words by frequency for sense tagging. We divided the Hansard vocabulary into ten groups based on frequency. The words in the first group were selected to be more frequent than words in the second, which were more frequent than those in the third, and so on. Each of the ten groups contained approximately the same number of *tokens*. Thus, the first group consisted of a small handful of high frequency words, whereas the tenth group consisted of a large number



**Figure 1: Singulars and Plurals are Used Differently.** The horizontal axis shows the probability that a word will be a noun, rather than a verb or adjective, when seen in its base form (no added *s*). The vertical axis shows the probability that the same word will be a noun when it ends with the suffix *-s*. For example, in this study *brief* is rarely a noun (< 25%), while *briefs* is almost always a noun (> 95%). The correlation is .67, showing that the singular and plural forms are often used quite differently.

**Table 6: Low Frequency Words  
Discriminate Best**

Group	Accuracy	Min. Freq.
1	.50	985,989
2	.57	587,847
3	.62	287,033
4	.67	134,531
5	.67	70,817
6	.73	24,309
7	.80	8,455
8	.84	3090
9	.88	802
10	.88	1

of low frequency words. We were interested in seeing how word performance would vary with word frequency.

Table 6 (above) shows the performance on the sense tagging application described earlier, averaged over the six polysemous words in Table 2. The third column shows the cut points defining the non-overlapping groups (a partition of the words) as the minimum frequency for the group. The frequencies are absolute frequencies out of a corpus with about 20 million words. Table 6 shows very clearly that the lower frequency groups out-perform the higher frequency groups.

Table 6 shows that the most desirable single groups are those with the most words. Thus restricting vocabulary by frequency will not reduce dimensionality. We also investigated the marginal utility of each frequency class, that is, the difference in accuracy due to adding one group to a model already based on all groups of lower frequency, beginning with a model based only on the group with the lowest frequency words. Each subsequent class, including surprisingly group 1, increased the accuracy of the models.

A group at Bell Communications Research has been investigating the use of singular value decomposition (SVD) as a means of dimension reduction. Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) discuss its application in information retrieval. As they say, "A fundamental deficiency of current information retrieval methods is that the words searchers use are often not the same as those by which the information they seek has been indexed." Singular value decomposition of the term by document matrix (the matrix with documents as rows, terms as columns, and the count of each term in a given document as the entry) selects directions in which groups of terms are found as dimensions, thus helping to solve this indexing-word problem, as well as reducing dimensionality. The authors report improvement in performance.

However, this is not yet an automatic solution to high dimension discrimination problems, because the size of the term by document matrix easily becomes unmanageable. The work cited used up to 2000 documents with up to 7000 terms. The previous discussion suggests that arbitrary reductions from 100,000 terms to 7000 terms will pay a price in accuracy. And while 2000 documents may be useful for some specialized

collections, this is a small number when compared to the number of scientific papers published annually, say.

## 11 Significant Correlations are not Perfect Correlations

Our models, and most previous statistical models have assumed that the conditional correlations between terms are zero (the Bayesian coefficients are directly related to correlations between words and the selecting condition). In this section, “correlation” should be read as “conditional correlation.” The most manageable alternative, used in stemming, is to assume that the correlation is perfect. The fundamental problem in considering interactions or correlations between the 100,000 dimensions of these models is that while we can manage  $V = 10^5$  reasonably complex calculations, and have enough data to support the calculations, we have neither data nor disk space nor computer time to handle  $V^2 = 10^{10}$  calculations. On the other hand, the interactions might be important for a number of our applications because there are many significant positive correlations among various pairs of words in the vocabulary.

Mosteller and Wallace describe a theoretical model for adjusting a pair of words to account for their correlation (1968, section 4.7). They conclude “For our *Federalist* data the differences observed for the several methods suggest modest adjustments to the log odds.” They make no adjustments to the log odds. Salton also discusses correlations, starting with his equation 10.18. He concludes “... not enough reliable term-occurrence information can be extracted from available document collections to produce improvements.”

We have examined the question of dependency briefly for the sense discrimination problem. We follow the theory presented by Mosteller and Wallace. The results of their theory are as follows. Let the log odds due to word  $w_1$  be  $\gamma_1^2$ , the log odds due to word  $w_2$  alone be  $\gamma_2^2$ , and the correlation between the occurrences of  $w_1$  and  $w_2$  be  $\rho$ , and suppose without loss of generality that  $\gamma_1 > \gamma_2$ . Then if we knew just the evidence from  $w_1$  we would have log odds of  $\gamma_1^2$ . The additional evidence from  $w_2$  is

$$\frac{(\gamma_2 - \rho\gamma_1)^2}{1 - \rho^2}$$

Notice that for  $\rho = 0$ , the contribution is  $\gamma_2^2$ , as we currently assume. The derivation imposes a limit of  $\gamma_2/\gamma_1$  for  $\rho$ . At this limit, the contribution from the second word is zero.

From the two models for duty, we selected the fifty most important words (maximum of score difference times frequency in the training sample), and calculated all pairwise correlations. A few were striking:

Table 7: Large Correlations		
0.36	tablets	tagamet
0.31	illicit	trafficking
0.28	organized	crime
0.24	prescription	patent
0.21	prescription	prices
0.21	tablets	valium

However, these were the only pairs exceeding .2 in correlation. Another 26 pairs exceeded .1 in correlation; none were below -0.1. The mean of the correlations was .01; the standard deviation of the mean was .001. In short, we agree with Mosteller and Wallace: the effects of correlations are modest.

## 12 Summary

We have discussed three major discrimination problems with large (100,000) dimensional spaces: sense discrimination, information retrieval, and author identification. The large number of dimensions results in each case from the number of terms each of whose frequencies will vary by context. We also gave two examples, capitalization restoration and person/place discrimination, from a much larger class of specific discrimination problems in high dimensional spaces.

Methods for these high dimensional spaces can basically be divided into two types: "direct," dealing with all the dimensions, or "indirect," attempting to reduce the number of dimensions first. We have shown that a Bayesian log odds model is a useful direct tool for each of the problems cited. It may not be the best tool for any of them, after thorough study, but it is easy to apply and gives results competitive with those of other methods where such other methods exist. It therefore appears to be a useful first cut tool for high dimensional discrimination problems.

There are problems with these methods. The Bayesian models have a problem of overstating their evidence because positive correlations between dimensions (words) cannot yet be accounted for. Brute force approaches to accounting for these correlations are not feasible, so some heuristics are needed. Also, as one example in author identification showed, the methods are currently not robust: evidence from one word can swamp the bulk of evidence from all other words. The methods need to be developed to overcome this undesirable feature, and used with caution until then.

While indirect methods have been necessary until recently due to computer limitations, the approaches have been heuristic. Since it has not been possible until recently to compare indirect methods with direct methods, little is actually known about the indirect methods. What little is known suggests they will be of limited utility or limited generality.

## Acknowledgments

Discussions with Collin Mallows were vital to the development of the method presented here for the estimation of conditional probabilities. This paper has appeared in the *Annals of Operations Research*.

## References

- [1] Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer, "Word Sense Disambiguation Using Statistical Methods," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 264-270, 1991.

- [2] Church, K.W., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Glasgow, 1989.
- [3] Dagan, I., A. Ital and U. Schwall, "Two Languages are more Informative than One," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 130-137, 1991.
- [4] Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, 41, 1990.
- [5] Dempster, A., N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society (B)*, 39, 1977, pp. 1-38.
- [6] DeRose, S. "Grammatical Category Disambiguation by Statistical Optimization," *Computational Linguistics*, 14, 1, 1988.
- [7] Francis, W., and H. Kučera *Frequency Analysis of English Usage*, Houghton Mifflin Company, Boston, 1982.
- [8] Gale, W., and K. Church, "Estimation Procedures for Language Context: Poor Estimates are Worse than None," pp. 69-74 in *Proceedings in Computational Statistics*, 1990, K. Momirović and V. Mildner, eds., Physica-Verlag, Heidelberg, 1990.
- [9] Gale, W., and K. Church "A Program for Aligning Sentences in Bilingual Corpora," *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991a, pp 177-184.
- [10] Gale, W. and K. Church "Identifying Word Correspondences in Parallel Texts," *Proceedings of the DARPA Conference on Speech and Natural Language*, 1991b.
- [11] Gale, W., K. Church, and D. Yarowsky "A Method for Disambiguating Word Senses in a Large Corpus," *Computers and the Humanities*, 1993.
- [12] Harman, D., "How Effective is Suffixing?" *Journal of the American Society for Information Science*, 42, 1991, pp. 7-15.
- [13] Harris, Z., *Mathematical Structures of Language*, Wiley, New York, 1968.
- [14] Hearst, M., "Noun Homograph Disambiguation Using Local Context in Large Text Corpora," *Using Corpora*, University of Waterloo, Waterloo, Ontario, 1991.
- [15] Leacock, C., G. Miller, T. Towel and E. Voorhees, "Comparative Study of Statistical Methods for Sense Resolution," *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.
- [16] Merialdo, B., 'Tagging Text with a Probabilistic Model,' *Proceedings of the IBM Natural Language ITL*, Paris, France, 1990, pp. 161-172.

- [17] Mosteller, Frederick, and David Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, MA, 1964.
- [18] Salton, G., *Automatic Text Processing*, Addison-Wesley, Reading, MA, 1989.
- [19] Salton, G. and C. Yang, "On the Specification of Term Values in Automatic Indexing," *Journal of Documentation*, 29, 1973, pp. 351-372.
- [20] Yarowsky, D., "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," *Proceedings of COLING-92*, Nantes, France, 1992.
- [21] Yule, G. U., *Statistical Studies of Literary Vocabulary*, Cambridge University Press, Cambridge, England, 1944.
- [22] Zernik, U., "Tagging Word Senses in a Corpus: The Needle in the Haystack Revisited," *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, P. Jacobs, ed., Lawrence Erlbaum, Hillsdale, NJ, 1992.
- [23] Zipf, G. K., *Selected Studies of the Principle of Relative Frequency in Language*, Harvard University Press, Cambridge, MA, 1932.

# Acquisition and Exploitation of Textual Resources for NLP

Susan Armstrong-Warwick

ISSCO

University of Geneva

*e-mail:* susan@divsun.unige.ch

## Abstract

Electronic access to large collections of texts and their translations provides a new resource for language analysis and translation studies. Empirical and statistical methods offer the means to organize the data and develop alternative models in view of a better understanding of our use of language. From a practical point of view they provide a basis for progress in the performance of NLP systems. A prerequisite for this work is the availability of machine-readable texts in an appropriate format. This paper will present current initiatives to acquire and prepare the necessary textual resource for corpus-based work and review current methods under development to exploit the data.

## 1 Background

For a growing number of researchers, electronic access to large collections of texts and their translations has become an essential resource for language analysis and translation studies. Empirical and statistical methods are being developed to organize the data in order to elaborate more adequate models of the structure and use of natural languages. Reliable methods for English are now available to tag texts for part-of-speech, predict word sequences, recognize collocations and automatically align sentences with their translations. These methods offer a starting point for deeper studies and practical applications in varying fields such as lexicography, speech recognition and machine(-assisted) translation.

This quite recent and growing interest in corpus-based studies is somewhat reminiscent of the empirical and statistical methods popular in the 50s. Initial work on machine translation (MT) – one of the first computational linguistic applications – was then related to problems of code-breaking (Weaver 1949). However, the computing resources were far from adequate and the textual resources, necessary as a basis for the statistical models, did not exist. Technological advances in computing power have certainly favored the reintroduction of this approach, as has the growing availability of electronic texts.

Another important factor which has contributed to the interest in data-oriented methods is the realization that rule-based systems have not produced the desired results nor the hoped-for basis for future progress. However, this new direction has also brought forth

a number of critics, a debate which can be characterized in terms of ‘statistics-based vs. rule-based’, ‘empiricist vs. rationalist’ or ‘Shannon-inspired vs. Chomsky-inspired’. Ironically, this controversy was also very much discussed in the 50s when Bar-Hillel (1951), arguing for the importance of semantics, wrote:

Let me warn in general against overestimating the impact of statistical information on the problem of MT and related questions. I believe that this overestimation is a remnant of the time, some ten years ago, when many people thought that the statistical theory of communication would solve many, if not all, of the problems of communication. (p.172)

This very same communication model has, in fact, been revived in current MT work (Brown et al. 1988) and with it, the same debate. However, most researchers engaged in corpus-based studies do not regard statistics as a total solution to the problem of describing language but rather as a means of modeling what occurs in texts as part of different computational applications. One important direction in the field is to integrate probabilistic models into rule-based systems and conversely, to augment statistical models of language with more traditional linguistic information.

Based on recent work, it has become clear that the new data-oriented methods offer potential solutions to key problems in computational linguistics:

- **acquisition:** identifying and coding all of the necessary information
- **coverage:** accounting for all of the phenomena in a specific domain, a given collection of texts, an application, etc.
- **robustness:** accommodating ‘real data’ that may be corrupt, ungrammatical or simply not accounted for in the model
- **extensibility:** applying the model and data to a new domain, a new set of texts, a new problem, etc.

While these problems are not new, access to large text resources does offer the means to investigate new directions that promise some important progress in the field theoretically and also to help solve very practical problems such as building dictionaries, classifying proper names and unknown words, and identifying noun phrases and other collocations.

In what follows we will present some general issues in acquiring and preparing corpora and report on a number of data collection activities currently in progress in Europe and North America. We will then review a number of studies which have shown how this data can be exploited.

## 2 Availability of Textual Data

In the latter half of the 1980s, when interest in statistical methods and corpus-based work was emerging in the computational linguistic community, there was very little material widely available for research purposes. This situation is in contrast to the

speech community where probabilistic models and statistical methods had become the standard and where data gathering was thus considered an integral part of any project (cf. Church and Mercer (1993) for a discussion of the development of statistical methods in speech research and its effect on work in computational linguistics and Liberman (1992) for some case studies in how publicly available corpora have benefited the speech community).

Older corpora for English, such as the Brown Corpus (see Francis and Kučera 1982) were quite small by current standards and others such as the Birmingham Corpus (see Sinclair 1987) were not publicly accessible. In continental Europe, where the new interest in corpus-based studies has only recently emerged, the situation is similar: the texts held in the national language centers are either too expensive for the individual researcher, not accessible in a manner conducive for current methods<sup>1</sup> or simply not available to the public.

Though there is a vast potential amount of data in electronic form, little of this material is currently available to the research community. The texts are privately held in centers all over the world and the holders of the data are often printing houses that do not have ownership and distribution rights. This separation of holders and owners is also apparent in large organizations where the technical services managing the archives are quite separate from other departments. Simply identifying where the data is located is often a problem itself once the texts have been printed (and are thus no longer in use).

The lack of appropriate textual materials (in quantity and range of data) has restricted research work in various ways. For languages other than English very little material is available; thus work has concentrated on the English language and methods have been tailored to take advantage of some language specific phenomena, e.g., fixed word order and limited morphology. It remains to be seen how far these can be extended to other languages. Translation studies up to the present have concentrated on essentially one language pair and one text type due to the public availability of only one corpus.<sup>2</sup> The proprietary nature of much of the data currently in use has meant that work was often duplicated since sharing results was discouraged.

Fortunately, this situation is slowly changing and it is this progress we wish to document here. A number of initiatives (cf. below) have served to increase the awareness of the desire and need for public access to data and to demonstrate the interest in cooperating in the acquisition and preparation of these resources. The data that has been made available through initiatives and [some individual efforts] has tended to be a rather ad-hoc collection. The community has been working under the motto that almost any data are better than no data and certainly, the more, the better. However, once large amounts of texts do become available, the issue of how to construct a 'balanced' or 'representative' corpus will have to be addressed – what Walker (1991) has termed the "ecology of language".<sup>3</sup> In order to provide adequate coverage of language at a given time or for a given domain, we will need to consider matters such as style, register, text type, frequency, etc., Biber (1993).

---

<sup>1</sup>Many centers offer remote access through in-house query programs whereas current practices require that the entire text must be available for manipulation in one's own laboratory.

<sup>2</sup>The Canadian Parliamentary Debates referred to as the Hansard Corpus, available from the ACL/DCI.

<sup>3</sup>See Walker's paper in this volume. This topic is of central concern to all corpora developed for lexicographical work.

One important issue that any data collection enterprise must address is how to protect the interests of the originators of the texts – a matter of critical concern in this new electronic era. Whereas texts are normally acquired and consulted for their information and amusement value, the use of texts in corpus-based studies is quite different. The interest is in the use of language rather than the content of a given document. This view of texts measured in *kilobytes* rather than *content* is often difficult to explain to the data holder who views texts in terms of copyright issues, scientific, artistic or popular value or simply as a potential source of revenue. And unlike the past, when a research environment simply meant access to a well-endowed library (or inter-library loan system) and adequate computing resources, for corpus-based studies each research group must have a personal copy of all of the material.

In light of this situation, all data collection enterprises make formal agreements with the data providers and those who wish to use the data. In the case of the organizations described below, each applicant for data must sign an agreement not to redistribute the data and to respect all restrictions as stipulated by the data providers. Though many issues of access, copyright and data protection in general (e.g. sensitive or private material, text collections and derived data as a potential source of revenue, etc.) are still in need of clarification, these agreements provide the legal basis to guard against misuse.

### **3 Data Collection Initiatives**

We now turn to a brief description of the new text collection and distribution activities that have emerged over the past few years. We begin with the largely volunteer efforts and then look at the later official projects that will assure a sounder structural basis.

#### **3.1 ACL/Data Collection Initiative**

The first such initiative, the *ACL Data Collection Initiative (ACL/DCI)* was established in 1989 by the Association for Computational Linguistics. The ACL provided the aegis of a not-for-profit scientific society to oversee the acquisition and preparation of a large text corpus to be made available for scientific research and without royalties. The acquisition work was carried out on a volunteer basis in a somewhat opportunistic manner, relying on availability rather than concerns of balance or representativeness. The clean-up and preparation (minimal SGML mark-up) of the material was done by a few individuals (Liberman 1989).

In 1991 the ACL/DCI produced and distributed its first CD-ROM and hundreds of sites are now working with this data. The disk contains over 600 Kb of mostly American English data and includes a large collection of newspaper articles from the Wall Street Journal, a dictionary of English donated by Collins Publishers and some grammatically annotated data from the Penn Treebank Project (Marcus et al. 1993), among others; a second CD-ROM is currently in preparation.

#### **3.2 European Corpus Initiative**

A similar initiative was established in 1991, the *European Corpus Initiative* (see Thompson 1992), to acquire a large multilingual corpus for research work in Europe. In partic-

ular, emphasis was put on gathering texts in languages other than English to provide the basis for researchers in all European countries to work on their own national language. An additional goal was to acquire a set of parallel texts (texts and their translations) in light of the importance of multilingual document production in Europe and the interest in translation studies.<sup>4</sup>

A large amount of data has now been collected for most European languages, with at least 5 million words of text for each of the major languages. A variety of parallel corpora have also been acquired from international organizations and Swiss banks (English, French, Spanish and English, French, German, respectively). The texts are currently being prepared and will be available on a CD-ROM by the end of the year.<sup>5</sup>

### **3.3 Establishing Text Repositories**

The two aforementioned initiatives have been singled out as exemplary for a new direction to meet the needs of researchers in the computational linguistic community. These volunteer efforts are now slowly being followed up by official projects which should establish a funding basis and the proper infrastructure for a longer term development of these resources.

#### **3.3.1 Linguistic Data Consortium**

In the United States the *Linguistic Data Consortium (LDC)* was established by the federal government in recognition of the necessity to followup the largely informal efforts with a sounder structural basis. Another concern was to provide the resources to all researchers, not just those in large and private laboratories (who already had access to in-house data and/or a budget to acquire and prepare private collections). The LDC was founded in 1992 with an initial start-up grant from the Advanced Research Projects Agency (ARPA) to acquire, prepare and distribute material for the research community (Liberman 1992). In less than one year the LDC has produced nearly 100 CD-ROMs and is actively working on acquiring a great deal more data. One of the major goals for the next year will be the acquisition of multilingual text to support machine translation and other activities.<sup>6</sup>

### **3.4 Multinational Efforts**

In Europe, where the multilingual environment poses special problems for centralized action in this field, work has begun on defining a framework for further actions regarding building up textual resources in Europe. An initial feasibility study was carried out under a project called the *Network for European Corpora (NERC)*. A follow-up project intended to establish the appropriate infrastructure for the collection of texts and the

---

<sup>4</sup>The work has been sponsored by the *European Chapter of the ACL (EACL)*, the *European Network in Language and Speech (ELSNET)*, the *Network for European Reference Corpora*, and the *Linguistic Data Consortium (LDC)*. Most of the work has been carried out at HCRC, Edinburgh and ISSCO, Geneva.

<sup>5</sup>The CD-ROM will be pressed by the LDC; distribution in Europe will be assured by ELSNET (email contact: elsnets@cogsci.edinburgh.ac.uk) and in the US by the LDC.

<sup>6</sup>Contact: The Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305; email: ldc@unagi.cis.upenn.edu

distribution of the data in Europe will begin next year under the CEC funded project *RELATOR*.<sup>7</sup>

Another project to begin this year (in follow-up to the *ECI*) is the collection and preparation of a large multilingual corpus (Thompson 1993). The corpus will consist of a set of comparable polylingual documents in at least six European languages (newspaper articles in the field of finance) and a multilingual parallel corpus in all nine languages (most likely drawn from the Official Publications of the European Community).<sup>8</sup>

## 4 Data Preparation

Aside from the basic problems of acquisition and negotiating rights for distribution, making the data useful often requires a good deal of effort to 'clean-up' and reformat it. Simply having data in electronic form is not necessarily sufficient, though in the future, as electronic document publishing evolves and mark-up and coding standards are established, this problem may disappear. Given the amount of time this work currently implies in any corpus collection activity – a situation that is likely to continue for perhaps a decade or more – it is not a task to be underestimated.

Older texts which were prepared uniquely for printing are usually stored on tapes in an undocumented and complex format and the correspondence between the logical structure of the text and the typographical structure is often not easy to establish. The large collection of texts from the United Nations, recently acquired by the LDC is but one example of this (Graff 1993). The documents for English, French and Spanish were archived on tapes made by the Wang computer system, an efficient means for storage but not for automatic extraction of all the files. Extracting the actual text data from the tapes required considerable effort (with help from Wang itself) to decipher the system specific character coding, format control codes and file structure. The recovery of the parallel texts could only be done semi-automatically due to the somewhat ad-hoc filename conventions coupled with numerous human-introduced errors. These problems arise from the fact that designers of older systems did not foresee such an application: the mark-up language was developed for physical display purposes only, rather than for logical representation of the information.

Beyond concerns of text mark-up schemes for formatting of texts is the issue of standards for annotation, i.e., additional, interpretive mark-up added to the data. Given the relatively little amount of data widely available and the current explorations of what information can reliably be identified in texts, it is not surprising that each corpus project has adopted in-house conventions for, e.g., sentence and word marking, part-of-speech-tagging and phrasal bracketing. As long as the mark-up is clear, well-documented, unambiguous and easy to convert for local machine processing, different conventions may suffice for these tasks.

However, as the information associated with the data becomes more complex, the standards, or conventions, adopted do become an issue. One major international project,

---

<sup>7</sup>The two projects NERC and RELATOR are carried out under contract to the CEC, DG-XIII, Luxembourg; contact: Roberto Cencioni, Jean Monnet Bldg., 2920 Luxembourg, or Nino Varile, email: M444@eurokom.ie.

<sup>8</sup>The project is part of the CEC International Scientific Cooperation Program.

the *Text Encoding Initiative*, has been working on a set of guidelines for coding material for all types of text mark-up with special attention to the complex needs of humanities researchers.<sup>9</sup>

Working with texts in a multilingual environment also raises a number of issues not necessarily apparent when only working with one language. The interpretation of a given symbol may be different for a given language (e.g., alphabet code conventions). The problem is more serious when some information associated with textual data in one language does not have an equivalent in another language. Studies are currently under way to determine to what extent the essentially English-based tagging systems in use can be adopted to European languages which display a wider range of morpho-syntactic phenomena (Monachini and Östling 1992).

As more hand-corrected data are prepared with sophisticated linguistic mark-up, an expensive and time-consuming task, annotation standards become an issue. To promote the sharing of resources and comparison of results, common coding schemes become a desirable goal. One major European project, *MULTEXT*, which plans to make a large multilingual, (partially) hand-validated corpus available with annotations for logical text structure, sentence marking, tagging and alignment of parallel texts, will address this issue in a systematic way.

These few remarks on text preparation and mark-up point to a large range of issues that will have to be confronted as more data becomes available in a wide variety of languages. Low-level issues of character sets and text formatting codes are in need of standardization to enhance international exchange of data. For higher-level mark-up it is perhaps premature to look for any standardization in the field. As new methods evolve and are applied to the data and as these results are shared, new conventions and standards will certainly emerge.

In the remainder of this paper we will review the various corpus-based studies currently under development.

## 5 Exploiting the Data

The increasing range of new methods being developed to exploit the data can be followed in the rise in publications and the number of tutorials and workshops dedicated to this topic. Whereas in the 80s, a large proportion of research work in computational linguistics concentrated on improving (unification-based) grammar formalisms and extracting data from machine-readable dictionaries, the publications of the 90s are witness to the new interest in data-oriented approaches. The journal *Computational Linguistics*, for example, recently devoted a large two volume special issue to "Using Large Corpora" and the theme of the conference on Theoretical and Methodological Issues in Machine Translation '92 was 'empirical vs. rationalist methods'. Workshops and tutorials addressing these topics are now held regularly in conjunction with the major conferences on NLP.<sup>10</sup> This approach has also become the main focus of all work under the ARPA

---

<sup>9</sup> An initial set of guidelines was published in 1991, a more comprehensive version will be available this year. Contact: tei-l@uicvm.bitnet

<sup>10</sup> In the following sections we can only mention a few of the numerous studies currently underway. The interested reader is referred to the collections given in the references to the cited papers. Cf. the tutorial by Liberman and Schabes (1993) for a topical bibliography.

program.<sup>11</sup>

## 5.1 Part-of-Speech Tagging

In contrast to the in-depth studies of grammatical phenomena in limited domains, data-oriented methods focus on (statistically) easily observable phenomena that can be determined with certain reliability over large quantities of data. The new methods aim at total (though perhaps superficial) coverage. The most-well established of these methods is that of part-of-speech tagging (e.g., Church 1988 and Cutting et al. 1992<sup>12</sup>). Given a sequence of words as input, these programs assign a sequence of part-of-speech tags with a very high success rate. The programs consist of a lexical component to assign a set of potential tags to each word and a component to disambiguate over sequences of tags (based on n-gram models that compute the probability of a tag given a previous sequence of tags). These taggers serve as the basis for a wide range of subsequent tasks, e.g., as a pre-processor for a parser (cf. Hindle and Rooth 1993, Marcus et al. 1993), as a basis for identifying phrasal expressions (Church 1988, Cutting et al. 1992, Smadja 1993), and in applications such as speech recognition, information retrieval and computational lexicography.

The widespread use of taggers is due to their ability to work on large amounts of quite variable data (given an appropriate training phase). They also represent the first step in solving at least one aspect of the ambiguity problem, one of the major problems of natural language analysis.

## 5.2 Grammar Development

In recognition of the meager results that traditional grammar development has generally produced, efforts have turned to incorporating data-oriented methods to improve performance and coverage along two different lines. One approach is concerned with augmenting existing grammars and traditional methods with probabilities, the other with inducing new grammars from large corpora. What both have in common is the need for annotated material in the training process.

The inclusion of probabilities in a parser, by ranking the rules according to their frequency of use for a given corpus, is reported on in Briscoe and Carroll (1993). They argue for the need to accommodate linguistically motivated constraints in contrast with some of the grammar learning programs that assign regular but arbitrary structures to the texts. Similarly, Black et al. (1993) discuss the development of history based grammars meant to accommodate a large variety of information, proposing a division of the parsing problem "into two sub-problems: one of grammar coverage for the grammarian to address and the other of statistical modeling to increase the probability of picking the correct parse of a sentence" (p. 36). The work by Hindle and Rooth (1993) to determine correct attachment of prepositional phrases by lexical probabilities is an example of this view.

---

<sup>11</sup>Cf. the *Proceedings of the Speech and Natural Language Workshop*, published by Morgan-Kaufmann, which provide a rich source of information about current work.

<sup>12</sup>The latter program is available via anonymous ftp from parcftp.xerox.com. See the bibliography in Liberman and Schabes (1993) for references to the numerous taggers.

The second direction in grammar development from corpora has concentrated on methods that are reliable and efficient to induce grammars automatically from texts. Pereira and Schabes (1992) demonstrate how the inside-outside algorithm can be successfully used to infer the parameters of a stochastic context-free grammar from a partially bracketed corpus. Bod (1993) derives a grammar from a corpus of labeled bracketings using statistical techniques. A less computationally intensive approach is presented in Brill (1993), who relies on only a very small training corpus to induce a grammar by simple transformations, i.e., by adding and deleting parentheses.

It is perhaps worth noting that all of the work reported on above was only possible due to the availability of annotated corpora. In fact, most of the current projects used material prepared by the Penn Treebank (Marcus et al. 1993). There is also work underway on training grammars from unlabeled texts (see Kupiec and Maxwell 1992). The underlying idea is to probabilistically identify word equivalence classes for subsequent use in part-of-speech tagging and parsing programs.

### 5.3 Lexical Acquisition

One of the major bottlenecks in NLP development has been the human labor-intensive task of acquiring the necessary lexical resources. The efforts to re-use existing machine-readable dictionaries have only partially alleviated this problem, and for languages other than English, there are no dictionaries that contain the explicit and detailed subcategorization information as found in the popular learners' dictionaries. Methods are being explored to automatically derive subcategorization frames, identify syntactic and semantic classes, discover phrasal expressions and build bilingual dictionaries (cf. papers in the *Proceedings of the SIGLEX Workshop*, Boguraev and Pustejovsky, 1993).

In partial answer to the need for detailed syntactic information, Brent (1993) developed a program to identify subcategorization frames of verbs based on the occurrence of pronouns. Manning (1993) and Ushioda et al. (1993) also report on work to acquire subcategorization frames. In the program developed by Manning (1993), the tagged data are first run through a finite state parser to identify potential complements and then filtered on the basis of statistical regularities over the candidate words. Methods for proper name identification and classification, an important phenomena in texts of all kinds, have been developed by McDonald (1993) among others. Weischedel et al. (1993) discuss a range of probabilistic methods for identifying unknown words and for dealing with ambiguity in more robust NLP applications. Work on the identification of noun phrases and collocations, another major problem for current NLP applications, is reported on in Smadja (1993) and Kupiec (1993).

Access to textual data provides the resource for learning about the different uses of a word, in particular uses not previously attested to in dictionaries or simply overlooked by human introspection (Church and Hanks 1990). Class-based approaches to lexical discovery have been investigated by Futrelle and Gauch (1993) who automatically identify classes on the basis of mutual information and position, and Resnik (1992), whose work is also based on mutual information measures, the initial word classes being constrained by a thesaurus (WordNet) (Miller et al. 1990). Word associations as they occur in text are compared to psycholinguistic studies in Wettler and Rapp (1993). Pustejovsky et al. (1993) and Waterman (1993) demonstrate how lexical semantic information can be

identified in texts.

## 5.4 Work with Multilingual Corpora

With the growing availability of large amounts of parallel texts,<sup>13</sup> corpus-based studies on translation have begun to emerge. Reliable alignment techniques for different text types have been developed (Brown et al. 1991; Kay and Röscheisen 1993; Church 1993) with a high level of accuracy. These alignment methods work with very simple notions of similarity of patterns of sentence lengths and regularity of (approximate) word pairs across the texts. Extensions to refine problematic alignment cases have been proposed using cognates (Simard et al. 1992) and predefined word lists such as bilingual dictionaries and terminology banks (Catizone et al. 1989). Two more recent studies by Church (1993) and Chen (1993) allow for more robust alignment in case of corrupted data (e.g., misplaced footnotes or missing segments of texts).

Partially annotated multilingual data is being used in studies to automatically identify word pair correspondences Dagan et al. (1993), in word-sense disambiguation Church (1991), in example-based machine translation Sato and Nagao 1990; Matsumoto et al. 1993; and Sumita and Iida 1992) and even in fully automatic MT (Brown et al. 1988)<sup>14</sup>. Example-based machine translation, first advocated by Nagao (1984), relies on a database of structured bilingual texts which are automatically matched according to lexical and structural regularities and various distance measures based on, e.g. thesauri.

Intelligent access to multilingual texts also provides the basis for a new generation of tools for translators (des Tombe and Armstrong 1993; Shemtov 1993; Simard et al. 1992). These new systems provide access to previously translated texts as a resource for identifying possible translations by searching on aligned text segments. The tools can also provide facilities for checking for potential translation errors such as missing segments and inconsistent use of terminology. Lexicography is another application domain where useful tools are being developed (Church and Hanks 1990) both for monolingual and bilingual work. In Smadja (1992), his initial work on extracting collocations is extended to include phrasal expressions and their translations. These multilingual investigations will certainly become more widespread as more parallel data becomes available.

## 5.5 Evaluation of Methods

An important issue which has been systematically addressed in US government funded NLP projects under the ARPA programs is the evaluation of methods and measurement of overall progress in the field. In Europe work is under way to elaborate policies and programs to better evaluate current work.<sup>15</sup> The issue of comparing results has been hampered by the limited textual resources available. The lack of public corpora for languages other than English has meant that much of the current work carried out

---

<sup>13</sup>Though currently only the Hansard corpus is publicly available a number of new corpora are currently in preparation by the LDC and the ECI.

<sup>14</sup>Statistical machine translation is an important focus of the ARPA program.

<sup>15</sup>E.g. under the recently created evaluation sub-group within the Expert Advisory Group for Linguistic Engineering Standards.

in different countries, working with different languages, has remained a local matter. And lack of comparable corpora in different languages has meant that no comparison is possible on how successful the current methods might be for languages other than English.

The issue on the adequacy of methods in use is yet another topic that deserves more attention in light of the potential misuse of statistical data. Church and Mercer (1993) address this issue in general and Dunning (1993) provides a case study of the potential weakness of using the wrong measures for a given problem. A comparison of different methods in view of a more systematic elaboration of evaluation techniques is presented in Grefenstette (1993) – a topic that will certainly become more important as methods proliferate and the ‘claimed’ results are brought under more rigorous scrutiny.

## 6 Conclusion

In this paper we have summarized a new and exciting direction in work in NLP. The growing availability of on-line corpora provides the basis for development of new methods to account for natural language phenomena, to further our insights in language use and to develop practical NLP programs. The necessary textual resources are still lacking, but some progress has been made to overcome this problem and current programs promise to deliver even more in the future. A representative sample of the wide range of new studies currently underway have been presented as a demonstration of the potential of the new data-oriented approaches to language study.

### Acknowledgments

The data collection work undertaken at ISSCO and reported on here has been supported by SWISSTRA and in part by a grant from the Linguistic Data Consortium. Thanks are also due to my colleagues Afzal Ballim, Graham Russell and Louis des Tombe for comments on previous drafts of this paper.

## References

- [1] Bar-Hillel, Y., “The state of machine translation in 1951”, *American Documentation*, 2:229–237, 1951.
- [2] Biber, D., “Using register-diversified corpora for general language studies”, *Computational Linguistics*, 19(2):219–242, 1993.
- [3] Black, E., F. Jelinek, J. Lafferty, M. Magerman, R. Mercer, and S. Roukos, “Towards history-based grammars: Using richer models for probabilistic parsing”, In *Proceedings of the ACL*, pages 31–37, Columbus, Ohio, 1993.
- [4] Bod, R., “Using an annotated corpus as a stochastic grammar”, In *Proceedings of the Conference of the European Chapter of ACL*, pages 37–44, Utrecht, Holland, 1993.

- [5] Boguraev, B. and J. Pustejovsky, (eds.) *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*. Association for Computational Linguistics, Columbus, Ohio, 1993.
- [6] Brent, M., “From grammar to lexicon: Unsupervised learning of lexical syntax”, *Computational Linguistics*, 19(2):243–262, 1993.
- [7] Brill, E., “Automatic grammar induction and parsing free text: A transformation-based approach”, In *Proceedings of the ACL*, pages 259–265, Columbus, Ohio, 1993.
- [8] Briscoe, T. and J. Carroll, “Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars”, *Computational Linguistics*, 19(1):25–60, 1993.
- [9] Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin, “A statistical approach to language translation”, In *Proceedings of COLING-88*, pages 71–76, Budapest, 1988.
- [10] Brown, P., J. Lai, and R. Mercer. “Aligning sentences in parallel corpora”, In *Proceedings of the ACL*, pages 169–176, Berkeley, California, 1991.
- [11] Catizone, R., G. Russell, and S. Warwick-Armstrong, “Deriving translation data from bilingual texts”, in Zernik, (ed.), *Proceedings of the Lexical Acquisition Workshop*, Detroit, Michigan, 1989.
- [12] Chen, S., “Aligning sentences in bilingual corpora using lexical information”, in *Proceedings of the ACL*, pages 9–16, Columbus, Ohio, 1993.
- [13] Church, K. and P. Hanks. “Word association norms, mutual information, and lexicography”, *Computational Linguistics*, 16(1):22–29, 1990.
- [14] Church, K. and R. Mercer. “Introduction to the special issue on computational linguistics using large corpora”, *Computational Linguistics*, 19(1):1–24, 1993.
- [15] Church, K., “A stochastic parts program and noun phrase parser for unrestricted text”, in *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas, 1988.
- [16] Church, K., “Concordances for parallel text”, in *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62, Oxford, England, 1991.
- [17] Church, K., “Char\_align: A program for aligning parallel texts at the character level”, in *Proceedings of the ACL*, pages 1–8, Columbus, Ohio, 1993.
- [18] Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun, “A practical part-of-speech tagger”, in *Proceedings of the Conference on Applied Natural Language Processing Processing*, Trento, Italy, 1992.

- [19] Dagan, I., W. Gale, and K. Church. “Robust bilingual word alignment for machine aided translation”, in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8, Columbus Ohio, 1993.
- [20] des Tombe, L. and S. Armstrong, “Using function words to measure translation quality”, In *Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and Text Research*, pages 1–18, Oxford, England, 1993.
- [21] Dunning, T., “Accurate methods for the statistics of surprise and coincidence”, *Computational Linguistics*, (19)1:61–74, 1993.
- [22] Francis, W. and H. Kučera, *Frequency Analysis of English Usage*. Houghton Mifflin, Boston, Massachusetts, 1982.
- [23] Futrelle, R. and S. Gauch, “Experiments in syntactic and semantic classification and disambiguation using bootstrapping”, In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, pages 117–127, Association for Computational Linguistics, Columbus, Ohio, 1993.
- [24] Graff, D., “The UN multilingual text corpus”, in *LDC Newsletter*, Vol. 1, No. 3. Linguistic Data Consortium, 1993.
- [25] Grefenstette, G., “Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches”, In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, pages 128–142, Association for Computational Linguistics, Columbus, Ohio, 1993.
- [26] Hindle D. and M. Rooth. “Structural ambiguity and lexical relations”, *Computational Linguistics*, 19(1):103–120, 1993.
- [27] Kay, M. and M. R'oscheisen, “Text-translation alignment”, *Computational Linguistics*, 19(1):121–142, 1993.
- [28] Kupiec, J., “An algorithm for finding noun phrase correspondences in bilingual corpora”, in *Proceedings of ACL*, pages 17–22, Columbus, Ohio, 1993.
- [29] Kupiec, J. and J. Maxwell, “Training stochastic grammars from unlabelled text corpora”, in *Workshop Notes from the AAAI Workshop on Statistically-Based Natural Language Processing Techniques*, pages 14–19, San Jose, California, 1992.
- [30] Liberman, M. and Y. Schabes, “Tutorial on statistical methods in natural language processing”, held in conjunction with the Conference of the European Chapter of ACL, 1993.
- [31] Liberman, M., “Text on tap: The ACL/DCI”, in *Proceedings of the 1989 DARPA Speech and Natural Language Workshop*, Cape Cod, Massachusetts, 1989.
- [32] Liberman, M., “Introduction to the Linguistic Data Consortium”, distributed at COLING-92, Nantes, 1992.

- [33] Manning, C., “Automatic acquisition of a large subcategorization dictionary from corpora”, in *Proceedings of the ACL*, pages 235–242, Columbus, Ohio, 1993.
- [34] Marcus, M., B. Santorini, and M. Marcinkiewicz, “Building a large annotated corpus of English: The Penn Treebank”, *Computational Linguistics*, 19(2):313–331, 1993.
- [35] Matsumoto, Y., H. Ishimoto, and T. Utsuro, “Structural matching of parallel texts”, in *Proceedings of ACL*, pages 23–30, Columbus, Ohio, 1993.
- [36] McDonald, D., “Internal and external evidence in the identification and semantic categorization of proper names”, In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, pages 32–43, Association for Computational Linguistics, Columbus, Ohio, 1993.
- [37] Miller, G. et al., Five papers on WordNet, Technical report, Cognitive Science Laboratory, Princeton University, 1990.
- [38] Monachini, M. and A. Östling, “Morphosyntactic corpus annotation - a comparison of different schemes”, Technical report, Istituto di Lingistica Computazionale, CNR, Pisa, 1992. Report for NERC project.
- [39] Nagao, M., “A framework of a mechanical translation between Japanese and English by analogy principle”, in A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence*, pages 173–180. North-Holland, 1984.
- [40] Pereira, F. and Y. Schabes, “Inside-outside reestimation from partially bracketed corpora”, in *Proceedings of ACL*, pages 128–135, Newark, Delaware, 1992.
- [41] Pustejovsky, J., S. Bergler, and P. Anick, “Lexical semantic techniques for corpus analysis”, *Computational Linguistics*, 19(2):331–358, 1993.
- [42] Resnik, P., “WordNet and distributional analysis: a class-based approach to lexical discovery”, in *Workshop Notes from the AAAI Workshop on Statistically-Based Natural Language Processing Techniques*, pages 54–64, San Jose, California, July, 1992.
- [43] Sato, S. and M. Nagao, “Towards memory-based machine translation”, in *Proceedings of COLING-90*, pages 247–252, Helsinki, 1990.
- [44] Shemtov, H., “Text alignment in a tool for translating revised documents”, in *Proceedings of the European Chapter of the ACL*, pages 449–453, Utrecht, Holland, 1993.
- [45] Simard, M., G. Foster, and P. Isabelle, “Using cognates to align sentences in bilingual corpora”, in *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–82, Montreal, 1992.
- [46] Sinclair, J. (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, Collins, London, 1987.

- [47] Smadja, F., "How to compile a bilingual collocational lexicon automatically", in *Workshop notes from the AAAI Statistically-Based NLP Techniques Workshop*, pages 65–71, San Jose, California, July, 1992.
- [48] Smadja, F., "Retrieving collocations from text: Xtract", *Computational Linguistics*, 19(1):143–178, 1993.
- [49] Sumita, E. and H. Iida, "Example-based natural language processing techniques - a case study of machine translation", in *Workshop notes from the AAAI Statistically-Based NLP Techniques Workshop*, pages 90–97, San Jose, California, July, 1992.
- [50] Thompson, H., "European Corpus Initiative", *ELSNEWS*, 1(1), 1992.
- [51] Thompson, H., "Multilingual corpora for cooperation (MLCC)", Proposal submitted under the LRE program for International Scientific Cognitive science-operation, 1993.
- [52] Ushioda, A., D. Evans, T. Gibson, and A. Waibel, "The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora", in *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, pages 95–106, Association for Computational Linguistics, Columbus, Ohio, 1993.
- [53] Walker, D., "The ecology of language", in *Proceedings of the International Workshop on Electronic Dictionaries*, pages 1–22, Tokyo, Japan, 1991.
- [54] Waterman, S., "Structural methods for lexical/semantic patterns", in *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, pages 128–142, Association for Computational Linguistics, Columbus, Ohio, 1993.
- [55] Weaver, W., Translation. (memorandum), 1949.
- [56] Weischedel, R., M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci, "Coping with ambiguity and unknown words through probabilistic models", *Computational Linguistics*, 19(2):359–382, 1993.
- [57] Wettler, M. and R. Rapp, "Computation of word associations based on co-occurrences of words in large corpora", in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 84–93, Columbus Ohio, 1993.

# The Center for Electronic Texts in the Humanities

Susan Hockey

Center for Electronic Texts in the Humanities

Rutgers and Princeton Universities

e-mail: *hockey@zodiac.rutgers.edu*

## Abstract

The Center for Electronic Texts in the Humanities (CETH) was funded in 1993 by the National Endowment for the Humanities to act as a national focus for the scholarly use of electronic texts in the humanities within the USA. CETH has three major activities: an Inventory of Machine-Readable Texts in the Humanities which is held on RLIN, the establishment of focussed collections of scholarly texts for access over the Internet via appropriate software, and an annual summer seminar on methods and tools for electronic texts in the humanities. CETH also provides information and support services for these activities. A Consortium of member institutions will be established. Consortium members will work together with CETH in our long-term objective of conducting research on the 'uses and users' of electronic texts in the humanities and using the results of this research to provide a continuous enhancement path for our facilities.

## 1 Introduction

The Center for Electronic Texts in the Humanities (CETH) was established by Princeton and Rutgers Universities in 1991 to provide a national focus within North America for those who are involved in the creation, dissemination and use of electronic texts in the humanities. Electronic texts are increasingly becoming major scholarly resources, but it must be acknowledged that we do not yet have a clear understanding of their promise and potential for the scholar and teacher. CETH's mission is to take a leading role in formulating effective methodologies for developing, maintaining and using electronic texts created by individual scholars or projects, and to establish a framework for advancing scholarship in the humanities by the use of high quality electronic texts. CETH represents the first concerted endeavor in North America to bring the use of electronic texts into the center of the humanities scholarly arena by building on existing resources and know-how. CETH's aim is to bring together the skills and methodologies developed in humanities computing for using electronic texts, and the expertise in libraries for managing and organizing information and providing access to it.

The picture of humanities research and teaching using computers over the past thirty years is one of diverse projects and applications. Scholars have used electronic texts for applications including concordances and word frequency indexes for print publication, analyses of style and authorship based on vocabulary usage, the production of historical dictionaries, studies involving lexis and some simple forms of syntax and morphology,

and the preparation of scholarly editions. The texts that have been created for these applications have remained with individuals or research groups and have often not been made available for others to use. These texts have many different markup schemes and most were not designed for reusability. Apart from one or two exceptions, notably the Oxford Text Archive which began in 1976, there have been few recognized efforts to make existing electronic texts available for other scholars to use, or even to catalog their existence. By the early 1990's, we can see that much enterprising work has been done, but we have also come to realize that there are no recognized procedures or standards for providing access to the texts or for maintaining them for the long term.

In the last two or three years, libraries have begun to acquire full text electronic materials, but the tools they use for handling them are, for the most part, adapted from those used for static print material and are thus generally inadequate to deal with dynamic electronic material, particularly that as complex as many humanities-related source texts. Efforts so far to provide electronic texts for the humanities in the library environment have concentrated mostly on commercially-available products, particularly CDROMs which fit better into current library procedures for handling information. Most of these CDROMS are packaged with specific software. The texts can only be used with that software, which, while satisfying some scholarly needs, can also be restrictive in the facilities provided. With a few notable exceptions, e.g., the library of texts made available at the University of Virginia, libraries have not dealt with the very many texts which exist in ASCII form and which therefore require software programs to make them usable. Those who have indexed texts for access via PAT or other retrieval software have had to make choices in the index-building. Those choices determine the questions which scholars can ask, but so far there has been little research to determine whether these are the questions which scholars want to ask.

Thus there is a strong need to develop and evaluate new methodologies for research and teaching using electronic texts. CETH's long-term objective is research on the 'uses and users' of electronic texts in the humanities and to use the results of this research for continuous enhancement of our facilities. Our two major initial projects are an international Inventory of electronic texts in the humanities and the development of access methods to collections of electronic texts over the Internet. These two activities are our means of exploring the use of electronic resources in the humanities. Our development plans include the establishment of a Consortium of institutions which will subscribe to our services and form a testbed for research. From what we learn from this research, we will be able to identify more clearly what else is needed to provide high quality electronic scholarly resources in the humanities. As the Consortium develops, we will use the experiences and expertise of its members to refine CETH's objectives and to assist in determining priorities to achieve those objectives, and then to work together with the members of the Consortium to meet these needs.

## 2 History of CETH

For some time there had been discussions in the USA about a national center for computing in the humanities. In March 1990 Marianne Gaunt and Robert Hollander organized a planning meeting at Princeton for a center to support the Rutgers Inventory of Machine-Readable Texts in the Humanities and to develop other related activities.

Some fifty scholars, representing all aspects of humanities computing, discussed various approaches. There was strong affirmation of the importance of the Inventory and a firm consensus that the center should develop educational programs. There was extensive discussion of the value and difficulty of establishing a text archive in the light of technology in common use at that time, namely the distribution of texts on tape, rather than over the network. There was, however, a general willingness to explore ways of making texts more accessible. In 1991 the National Endowment for the Humanities (NEH) provided funds to support the Inventory for one year and until a director was in place, and the Andrew W. Mellon Foundation also contributed startup funds. CETH was established in late 1991, and in 1993 substantial support was awarded by the NEH for three years.

### 3 The Inventory

Given the amount of time that was, and still is, necessary to create an electronic text and the lack of any one catalog, it is not surprising that much effort can be expended on establishing whether a copy of a particular electronic text already exists and whether it can be used by other individuals. It was for this reason that the Rutgers Inventory of Machine-Readable Texts in the Humanities was begun in 1983. It is a catalog of existing electronic texts in the humanities and is held on the Research Libraries Information Network (RLIN). With support from various sources, the Inventory had grown to some 1600 entries in 1991 when CETH was established and took it over. Of these about half are entries from the Oxford Text Archive (OTA) cataloged with support from the Mellon Foundation in 1990-91.

Our initial estimates taken from sources such as the Humanities Computing Yearbook indicate that there are already tens of thousands of electronic texts in the humanities. The combined holdings alone of ARTFL (American Research on the Treasury of the French Language), the Thesaurus Linguae Graecae and the Institute of Computational Linguistics at Pisa number over 10,000. One catalog record is initially created for collections such as these, and texts within the collections will be cataloged individually at a later date.

There are several very obvious differences between cataloging print materials and computer files. Appropriate software is often needed to examine the text in order to derive information about it. Unlike most other catalog records on RLIN, very few of the items are owned by the cataloging institution (the Center), which has to acquire information about them from their holders who are mainly individuals and research institutes which have little or no experience in cataloging and often do not have adequate documentation.

Chapter 9 of the Anglo-American Cataloging Rules, second edition, 1988 Revision (AACR2R) deals with computer files, but the rules in it cover all kinds of computer files (programs, numeric data files, etc.), and are not especially suitable for electronic texts, particularly those which are electronic representations of material which already exists in print or manuscript form. For example, the rules specify a title screen, which most electronic texts do not have, since they are plain text files without software to display any titling information. The physical characteristics of a file are also less important because of the ease of copying from one medium to another, and we do not usually record this

information except in the case of CDROM. We do include the encoding scheme and whether the text needs specific software, since this information is necessary for any user of the text. In contrast to the static form of a book, an electronic text is a dynamic object which can be amended and updated many times. As the Inventory develops, there will be a need to establish procedures for determining what constitutes a new version and to amend existing records as new versions of those texts become available.

Research done at CETH shows that there is little expertise in the specific problems of cataloging electronic texts. Over the last year or so CETH's cataloger has developed extensive guidelines for cataloging electronic texts and we are now receiving requests from other institutions for them. These guidelines are based on our experience in cataloging CDROMs and other texts which the CETH owns as well as the responses to a major survey conducted in 1992-3 and discussions with other groups.

The survey was initially part of a European Community-funded project. It began in 1990 when a first questionnaire was distributed to some 5000 people who were on the mailing lists of the Association for Literary and Linguistic Computing, the Association for Computers and the Humanities, the Association for Computational Linguistics, and twelve other sponsoring organizations. Recipients were asked to note which of the following types of data they hold: (1) speech (tape recordings), (2) single (individual) texts, (3) collections of texts, (4) corpora (5) machine-readable dictionaries, (6) computational lexica. The intention was to send a different follow-up questionnaire depending on the type of data held. CETH took over responsibility for doing the follow-up for (2), (3) and (4).

The second follow-up questionnaire formed three parts. Part A requested information about the hardware and software environment used, Part B was concerned with single texts and Part C merged (3) and (4) above. Antonio Zampolli provided a draft questionnaire which was subsequently revised and redesigned with considerable assistance from Don Walker. In redesigning the questionnaire CETH took great care to ensure that the responses would be given in such a way that bibliographic records could easily be created, although we recognized that many of the respondents would not be familiar with the use of bibliographic records. Examples were provided for several different types of texts and a written explanation given. The questionnaires were sent out at the end of August 1992 to some 400 people who responded to the earlier questionnaire. At the time of writing, we are still going through the responses which have provided a lot of information. Some responses were very detailed, but others gave only outline information. It seems that the questionnaire was perhaps too lengthy and some redesign might be needed for the future. We have received much valuable assistance in evaluating the survey responses from Roberta Brody, a PhD candidate at Rutgers who has considerable experience in consumer intelligence for libraries. We will use her guidance in the design of future questionnaires.

In collaboration with the Oxford Text Archive we are also now investigating ways of using the Text Encoding Initiative's (TEI) proposals for an electronic document header to speed up the cataloging process. The TEI header at the beginning of a TEI electronic text file contains bibliographic information about the source text and the encoding used within it. The Oxford Text Archive now makes some texts with TEI headers available by ftp and we have been accessing these to determine how we can make best use of the headers.

## **4 The Text Collection**

The Text Collection forms the second major activity within the Center. We are only just now beginning this with the aid of the recent NEH funding. We wish to make texts available in the scholarly environment that have as much authority and integrity as printed and published works. CETH's acquisitions policy will be to concentrate on a number of focussed collections of material which are properly tagged and documented and which represent good scholarly editions. We envisage a time when it will be just as normal for scholars to use the network to access full text databases as they now do for bibliographic databases and electronic mail. They will use the Inventory as a first means of access and move seamlessly from that to the Text Collection at CETH or to collections of text stored on other computers on the network.

Concentrating on focussed collections will help to fulfill our mission of making usage of electronic texts central to humanities scholarship where users should expect the same quality of material as they find in a printed book or source, with adequate tools for accessing it. Up to now many electronic texts have been in the public domain and of variable quality. But recently there have been moves towards better policy, acceptability and control for electronic texts. Support for this view has been expressed for some time by leading practitioners in humanities computing, and librarians are now beginning to establish collection development policies for electronic texts. Together with our Consortium, CETH expects to play a leading role in establishing these guidelines which will be formulated to satisfy the needs of high quality scholarship. Making the material available through the library environment also implies permanence and standardization, which has been lacking in other attempts to collect and distribute electronic text.

Access via the network is the most effective way for texts to reach as wide an audience as possible with as much convenience and flexibility as possible. The planned development of the National Information Highway is now bringing this nearer to reality in the USA. Network access allows many more texts to be delivered, with the potential for using a broad range of different kinds of software which is essential for dealing with the complex intellectual problems of scholarly texts. It also provides CETH with better control over the texts and a means of updating them centrally without the need to send out new versions individually to many different users.

We plan a two-tiered method of access. The first will be for on-line searching. The second will permit downloading of texts for scholars to manipulate the texts, to serve their own needs. The first method of access will be important for classroom use and also for broad searches over a range of material. The second will enable scholars to use their own software, which is specific to their own individual requirements, and will thus permit greater flexibility in overall use. We consider maintaining the integrity of the texts to be crucial and will co-operate with other organizations that are also developing mechanisms for this purpose.

CETH has made a very strong commitment to the Text Encoding Initiative (TEI) implementation of the Standard Generalized Markup Language (SGML). With its 'descriptive' rather than 'prescriptive' approach to markup, SGML provides a mechanism which has a sound theoretical and intellectual basis, for describing the characteristics of a text. Besides the obvious reasons for standardization, encoding in SGML requires a thorough analysis of the intellectual issues involved in creating an electronic text. It

provides a much better mechanism than any other encoding scheme for handling the complexities of scholarly texts in the humanities, for example the critical apparatus, marginal notes, changes of language and script, for which adequate encoding schemes have never previously been available.

SGML is also application-independent. The same text may be used for many different purposes including printing, browsing, searching and building hypertextual links. Texts encoded in SGML represent an investment for the future. Their longevity can be assured as texts are transferred to new systems.

However SGML does have one short-coming. It is built on the assumption that the document being encoded is one single hierarchy in structure, but most humanities texts contain several hierarchies reflecting multiple views of the text. Although SGML does provide some facilities for multiple hierarchies, these are somewhat clumsy and not widely used.

The TEI has designed and developed an SGML application to serve the needs of humanities scholars. The TEI Guidelines give recommendations both on what features to encode and how to encode them, and thus include a comprehensive list of features likely to be found in humanities-related material. These features include both those which are explicitly marked and those which are the result of analysis and interpretation of the text. Although the TEI Guidelines include some 400 different encoding tags, very few indeed are absolutely required. The basic philosophy is 'if you want to encode this feature, do it this way'. Sufficient information is provided for the encoding to be extended by users if necessary.

The Guidelines are built on the assumption that virtually all texts share a common core of features, to which can be added tags for a specific discipline, text type or application. The encoding process is seen as incremental, so that additional tags may be inserted in a text as new scholars work on the text. Almost all encoding implies some interpretation of a text and so the Guidelines provide for multiple views of a text and multiple encodings for individual phenomena within a text. They also provide a means of documenting any interpretation so that a new user of the text can know why that interpretation is there. A further issue in encoding is fidelity to and recoverability of the source. Here the choice is left to the encoder who is again strongly encouraged to document the decisions made.

A user of the Guidelines selects one or more base tag sets which are relevant for the type of text (e.g., prose, verse, performance texts, dictionaries) to which are automatically added tags for features common to all texts and those needed for documenting the text. If appropriate, additional tag sets (e.g., for textual criticism, hypermedia, simple analysis, names and dates, etc.) may also be included. The TEI document type declaration (DTD) is thus built up using a model which has been likened to the preparation of a pizza.

All CETH texts will be maintained in the TEI format and will be validated against the TEI DTDs before being made available. They will have a full TEI header which can be used for cataloging as well as providing documentation about the encoding and other ancillary information which is necessary for scholarly use of the text. The TEI has also proposed various ways of dealing with multiple hierarchies in SGML and we expect to act as a testbed for some of these in order to test their applicability for different kinds of humanities material.

Commitment to SGML is now widespread and a number of software companies have

developed commercial systems for supporting SGML and its use is spreading further into the publishing and information retrieval markets. However, these software developments in most cases fall short of what is needed to satisfy the varied requirements of the humanities scholar. CETH has been conducting a more detailed analysis of software for on-line searching and for mechanisms which will enable texts to be accessed over the network. Besides retrieval, these will include hypertextual links, maintenance of the texts, monitoring of use, menu design and control of access. At the time of writing, a very preliminary report is available, which indicates that most existing systems fall rather short of some basic requirements for humanities scholarship. It seems that most existing SGML software does not handle attributes as well as we would like. The TEI makes extensive use of these for cross-references as well as for functions like joining together the parts of a 'dismembered quotation', defining the components of a name or date, or aligning parallel multilingual texts.

We would also like to use SGML tagging to improve text retrieval and to experiment with some of the tools and techniques which are more widely used in computational linguistics. Humanities texts are in general more complex than scientific and technical material and they can be used for many purposes other than the retrieval of content. To date, humanities computing has made little use of word class tagging or lexical database techniques. We will plan the Text Collection and access mechanisms with these developments in mind so that in the future we can experiment with providing more sophisticated intellectual access, once the basic procedures for using the Text Collection have been established.

The Text Collection is beginning with the Women Writers Project (WWP) from Brown University which is creating a textbase of all women's writing in English for the period 1350-1830. It is an ideal test database for CETH's text collection because it is a body of material in English which is scholarly in nature yet has potential wide usage both for research and instruction, not only in literature, but also in history, sociology and related disciplines. It has been designed for many potential uses from printing to searching and browsing, and performing comparative studies and it is encoded in TEI-conformant SGML.

We are also preparing to work with various collections of documents and papers of historical interest and have plans to co-ordinate some model projects in documentary editing. The rapid growth in interest in multimedia and developing potential for it means that these collections can be accompanied by photographs, manuscript images, sound and other non-textual material.

The Text Collection will serve as a test-bed for CETH's major objective of developing an effective methodology for working with electronic texts in the humanities. Once some texts are set up and start to be used, we will learn from providing access to them, and studying how they are used. We will thus be able to establish a basis for further research on the use and users of scholarly texts in the humanities. As we add further collections of text, we will be able to identify more closely those scholarly needs which are not well supported by computational tools and encourage the development and evaluation of new tools. Research projects which are supported within CETH or to which CETH is linked will provide testbeds within which software can be evaluated and new methodologies developed. As CETH develops, we plan to seek funding elsewhere for short-term fellowships and other initiatives which will bring people at CETH to work on existing

electronic text projects and develop new ones, using the equipment and facilities at CETH.

As a first step towards identifying better the needs of scholars, CETH is collaborating in research sponsored by Bellcore and conducted by Nicholas Belkin of the Rutgers School of Communication, Information and Library Studies. Professor Belkin is an information scientist who is researching design principles for using electronic resources. With Don Walker, he has been studying the ways in which scientists interact with full text retrieval systems. He has now been conducting interviews with some humanities scholars, initially classicists at Rutgers and Princeton, in order to elicit how their needs differ from those of scientists. He participated in a panel on this topic organized by CETH at the 1993 ACH-ALLC conference, which, as far as can be determined, was the first time that user needs have been discussed in this way at a humanities computing conference. This research is still at an early stage, but it is already providing some preliminary results and we expect to see it develop into a substantial activity.

## 5 Summer Seminars

CETH has now organized two highly successful summer seminars (in 1992 and 1993) with the co-sponsorship of the Centre for Computing in the Humanities (CCH), University of Toronto. For each seminar thirty participants were selected to spend two weeks at Princeton in intensive study of methods and tools for electronic texts in the humanities. The seminar addressed a wide range of challenges and opportunities that electronic texts and software offer to teachers, scholars, and librarians in the humanities. It was intended for faculty, students, librarians, technical advisers and academic administrators with direct responsibilities for humanities computing support. The focus of the seminar was practical and methodological with the immediate aim of assisting participants in their teaching, research and advising. It was concerned with the demonstrable benefits of using electronic texts, with typical problems and how to solve them and with the ways in which software fits or can adapted to common methods of textual study.

In the first seminar, the following topics and software were covered: what electronic texts are and where to find them, survey of existing resources, introduction to simple concordancing with MTAS, creating and capturing text in electronic form, demonstration of OCR, introduction to encoding, Micro-OCP, TACT, stylistic comparisons and introductory statistical tools, critical editions, methods of scholarly publishing, advanced tools (lemmatization, morphological analysis etc), electronic dictionaries, and some hypertext. The instructors were Willard McCarty of CCH and Susan Hockey. The second seminar covered similar ground, but, following comments from the first group of participants, included the additional topics of the TEI, taught by C.M. Sperberg-McQueen, TEI Editor, and a more in-depth look at hypertext taught by Elli Mylonas, Managing Editor of the Perseus Project.

Each seminar also included a number of special lectures. In the first year, Don Walker gave a global view of electronic resources for text and language analysis, and Elli Mylonas presented Perseus and Pandora. In both years Bob Hollander demonstrated the Dartmouth Dante Project and staff from Princeton showed ARTFL and the New Oxford English Dictionary. In the second year George Miller described and demonstrated

WordNet, and Paul Peters gave an overview of the work of the Coalition for Networked Information.

Participants were able to spend time working on their own projects with assistance from the instructors and staff of Computing and Information Technology at Princeton. All the software used for the seminar was available to them 24 hours per day either in the computing center (daytime) or the student accommodation where they stayed (for the rest of the time). In 1992 projects included Piers Plowman manuscripts, the language of Nabokov, Durkheim's works, the diary of Robert Knox, morphological analysis of Latin, and a study of the Book of Mormon. In 1993 they included the CURIA Irish Manuscripts Project, the Papers of Thomas Edison, critical edition of medieval French, Swahili-Polish lexicography, Sanskrit narrative prose, modern poetry, an electronic edition of W.B. Yeats and an archive of Australian aboriginal material.

The seminar is attracting growing interest and international participation. In both years, six of the participants were from outside North America and there were several more applications from the other side of the world. It seems that demand for such a program will continue to grow for some time and CETH has recently been considering possible avenues for expansion. The two-week seminar with thirty participants has worked very well. Participants have got to know each other well and have derived a lot of benefit from close interaction among themselves and with the instructors. All sessions have been plenary and participants were expected to attend every session. However the number of applicants for 1993 was almost double that for 1992. Judging by the number of inquiries which we have already received for 1994, it seems unlikely that we can satisfy all the demand with the current format and some other model may be needed in the future.

## 6 Information Services

CETH also provides information services which are essential support for our major activities. CETH's newsletter is published twice per year and is circulated to all on our mailing list (about 1700 names at the time of writing) as well as to all Faculty at Rutgers and Princeton. We have deliberately chosen to have a print publication so that we can reach new people who do not yet make regular use of the networks, but we are now also making the newsletter available on the Gophers at Princeton and Rutgers. The newsletter content is designed so that it can be used to answer typical queries. It concentrates on activities related to CETH and does not duplicate other newsletters. The newsletter is national and international in its focus, and it has supplements with information specific to Princeton and Rutgers for distribution locally.

CETH also has an electronic distribution list, [ceth@pucc.princeton.edu](mailto:ceth@pucc.princeton.edu). In order to avoid duplication with other discussion lists the CETH list is used only for disseminating information about CETH, which is submitted by CETH. We also use a private electronic mail list to maintain contact with 'alumni' of the summer seminar.

We have recently been responsible for establishing a new discussion group on electronic text centers within the Association of College and Research Libraries. This group was formally approved at the American Library Association meeting in June 1993 and now has its own electronic discussion group which is moderated by CETH's cataloger, Annelies Hoogcarspel. The electronic group attracted 600 participants within its first

three weeks and continues to grow. Its focus is on issues such as acquisitions, cataloging, public services, budgets, management, training and staff development, and on full-text files that are primarily monographic in nature.

Outreach to the use community has been a major objective in the first two years of CETH's existence. We have participated in the annual conferences of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, the American Library Association, the American Society for Information Science, the Modern Language Association and the American Historical Association as well as numerous individual events. Besides more general talks and meetings, we have begun to concentrate our outreach activities on those groups who are likely to make substantial use of CETH in the future.

## **7 CETH at Rutgers and Princeton**

CETH is also providing support for electronic texts and humanities computing at Rutgers and Princeton. At Rutgers we are setting up an Electronic Text Center in the Graduate Reading Room of the Alexander Library. This Center has some additional support from Rutgers and at the time of writing we are training three student assistants who will help users in the Center, covering the time from 1pm until 9pm on weekdays. New equipment is on order and we expect to be fully operational by December 1993. We have purchased most of the primary source full-text humanities material which is available commercially. This includes the New Oxford English Dictionary, Perseus, the WordCruncher Library of American Literature, Shakespeare and other Oxford University Press Electronic Texts, the Thesaurus Linguae Graecae, the PHI CDROM of Latin literature, the Global Jewish Database, Goethe, CDWORD, Past Masters philosophy texts, the Aquinas CDROM, ICAME CDROM, CETEDOC Library of Early Christian Fathers and some Hegel texts. We are also providing an OCR service with a Scanjet IIC scanner and Typereader software on a Macintosh. We plan to involve users of these facilities in the development of CETH and will, as far as is possible, discuss their needs with them and survey which resources have most use and for what purposes.

At Princeton we have a collection of resources on CDROM and diskette which are available for use in our office in the Firestone Library. Most of the resources which are in the Rutgers center are duplicated at Princeton, which also has the more extensive version of the Global Jewish Database which includes all the Responsa material. At both institutions we are assisting Faculty and Students with their projects and at Princeton we work closely with humanities computing staff at Computing and Information Technology.

We have also begun to work together with the School of Communication, Information and Library Studies at Rutgers in other respects. In summer 1993, Lisa Horowitz, an MLS candidate, spent 150 hours at CETH as field experience for part of her MLS course. Using her skills in French and Italian as well as technical writing, she carried out an in-depth study of ARTFL and the Dartmouth Dante Project and wrote introductory user guides for each of these databases. She was also able to attend part of the summer seminar and thus gained a useful insight into the challenges which electronic texts pose for the reference librarian. A second MLS candidate already has plans to do field experience

with us in the summer 1994, this time in English literature, and further student projects are likely.

## **8 People**

At present the Center has three full-time staff members. The Director is responsible for overall management of the Center and supervising all activities as well as fundraising and outreach. A Cataloger is responsible for overall development of the Inventory and interacting with compilers of texts. The Cataloger has developed CETH's procedures for cataloging electronic texts and is working closely with other groups looking at issues of cataloging electronic resources. A Library Associate handles our information services, which include newsletter, discussion list, and answering e-mail, as well as the applications and day-to-day organization of the summer seminar. We will shortly be advertising a fourth full-time position, a Text Systems Manager who will be responsible for setting up and managing the Text Collection and access to it. We expect to recruit somebody with substantial skills in UNIX and SGML. Our short-term expansion plans include provision for a second cataloger. We already have some student assistance and we expect to have more in future as our activities grow.

CETH has an international Advisory Board of humanities scholars, information professionals, publishers, computer scientists and network specialists. The principal role of the Advisory Board is to provide broad guidance on the activities of CETH. It meets once a year, normally at the end of October, at Princeton's house at Dunwalke, Bedminster, NJ. CETH's Governing Board consist of four members each from Princeton and Rutgers. The Governing Board oversees the local activities of CETH. It meets three times a year, alternating between Princeton and Rutgers and its membership represents the interests of libraries, computing and Faculty.

## **9 Future Plans**

The development of the Consortium is essential for the future financial security of CETH. The NEH funding which we have received for 1993-6 is intended to support that development and establish services for Consortium members. We expect the Consortium to be fully operational by the end of the grant period. Our long-term goal is for CETH to be internationally recognized as the place that sets the lead in bringing high quality use of electronic texts into the center of the scholarly arena in the humanities. The cataloging will continue on RLIN, which is the major on-line source of bibliographic information, but it is recognized that this cannot be accessed everywhere. A master database of the Inventory will be maintained by CETH, from which it will be possible to generate the information in other forms and in particular to make it available over the Internet. The Inventory will provide an interface to our own and other text collections, so that users can move seamlessly from the catalog to the text. We will build up further collections of texts and plan to work with digitized images as well as text. These texts and the Inventory will be maintained on CETH's own host computer and will be closely integrated. Some of the texts will contain more sophisticated tagging which can be used

for retrieval of linguistic analyses and literary interpretation, where the latter is deemed appropriate.

We also plan to continue to organize a summer seminar each year, focusing on the use of electronic texts. The syllabus will keep up with new developments whilst still concentrating on basic methodologies, and we expect people who attend the seminar to disseminate the information in their own institution. Our program of research is an investment for the future. We will continue to assess and evaluate new techniques for scholarly applications of electronic texts and conduct research projects which will improve the intellectual level of access to these texts.

## References

- [1] Burnard, L., "What is SGML and How Does It Help?", TEI document TEI ED W25, available from TEI fileserver tei-l@uicvm, 1991.
- [2] Calzolari, N., A. Zampolli, "Lexical Databases and Textual Corpora: A Trend of Convergence between Computational Linguistics and Literary and Linguistic Computing", in S. Hockey, N. Ide (eds), *Research in Humanities Computing*, Oxford University Press, Oxford, 1991, 273-307.
- [3] Hockey, S., "The ACH-ACL-ALLC Text Encoding Initiative: An Overview", TEI document TEI J16, available from TEI fileserver tei-l@uicvm, 1991, revised 1993.
- [4] Hockey, S. (forthcoming), "Developing Access to Electronic Texts in the Humanities", in L. Saunders (ed.), *The Evolving Virtual Library: Visions and Case Studies*, Meckler, Westport.
- [5] Hockey, S., D. Walker, "Developing Effective Resources for Research on Texts: Collecting Texts, Tagging Texts, Cataloging Texts, Using Texts and Putting Texts in Context", forthcoming in *Literary and Linguistic Computing*, 8(4), 1993.
- [6] Hoogcarspel, A., "Bibliographic Control of Electronic Texts", paper presented at *ACHALLC93 Conference*, Georgetown, June 1993.
- [7] Horowitz, L., "CETH as a Field Experience", *CETH Newsletter* 1(2), 1993, 8-10.
- [8] Sperberg-McQueen, C.M., L. Burnard, (eds), "Guidelines for the Encoding and Interchange of Machine-Readable Texts", draft version 1.1, TEI document P1, ACH-ACL-ALLC, Chicago and Oxford, 1990.

# Design Principles for Electronic Textual Resources: Investigating Users and Uses of Scholarly Information\*

Nicholas J. Belkin

School of Communication, Information and Library Studies

Rutgers University

*e-mail:* [belkin@zodiac.rutgers.edu](mailto:belkin@zodiac.rutgers.edu)

## Abstract

We describe a project whose goal is to develop a coherent set of principles for the design of electronic textual databases to support scholarly activity, in particular in the humanities. The focus of the project is a series of investigations which aim to understand: the tasks and goals of scholars in the humanities; the behaviors of such scholars in their interactions with texts of all types; the circumstances which lead them to engage in particular text-related behaviors; and their explicit uses of texts to accomplish their tasks and achieve their goals. We report here on the initial stages of our overall project, describing our methods and presenting preliminary results of a classification of scholarly tasks and goals, and of behaviors in interaction with texts and parts of texts, and also suggestions about the relationship between goals and information-seeking activities.

## 1 Introduction

The availability of a wide variety of texts<sup>1</sup> in electronic form is an increasing, and increasingly important factor in the working milieu of the community of scholars and scientists. There seem clearly to be obvious advantages to having texts available in such forms (and disadvantages, as well), but, to date, almost no attention has been paid to the potential uses of such textual resources, nor to how the existence of such resources fits with the general working environment of those for whom they are intended, nor to the influence that the uses and their context should, or could, have on the design of electronic textual resources. By design, we mean explicitly the contents, structure, organization and access mechanisms of such resources. Rather, emphasis has been placed primarily upon getting texts into electronic form, either to support some very specific project, or with a general concept that such a resource would somehow "be useful". Recently, some notable exceptions to this general inattention to scholars' working environments

---

\*A preliminary version of this paper was presented at the 1993 ACH/ALLC Conference in Washington, D.C. We thank Susan Hockey for the making that presentation possible, and the participants in our session for their invaluable comments.

<sup>1</sup>By this term, we include not only linguistic texts, but also texts incorporating all relevant media.

have begun to appear. Examples include *Humanists at Work* (1989), Bates (1993), Chu (1993) and Wiberley (1993).

Following Walker (1981), we propose that it is now time explicitly to consider just how those for whom texts, in whatever form or medium, are important in their work, actually use (or wish to use) them, in order to understand how best to design electronic textual resources which would provide more effective, or even new, means of support for such people in their working lives. In this paper, we describe a project whose goal is the establishment of principles for the design of truly effective electronic textual resources, based on understanding of why and how people use texts, or information. This is a collaborative project between the Rutgers School of Communication, Information & Library Studies, Bellcore, and the Rutgers/Princeton Center for Electronic Texts in the Humanities. It was begun under the impetus of Don Walker of Bellcore, and continues in his spirit.

## 2 The Overall Project Plan

In order to understand the roles of texts in the work of scholars and scientists, at least four issues need to be addressed, which we describe in some detail below. These issues form the basis for our investigative design.

First, one needs to understand the overall goals and tasks which motivate these people's activities in knowledge production. By understanding, we mean explicitly:

- discovering and describing the range and variety of such tasks;
- categorizing these tasks/goals;
- relating such categories to those developed by others in other domains and contexts.

Second, one needs to understand and characterize the behaviors that these people engage in in their interactions with texts (or, perhaps more generally, with information), and the intentions which lead from the tasks to these explicit behaviors. Here, by understanding, we mean:

- describing and categorizing the interaction behaviors; and,
- relating these behaviors, and the goals associated with them, to one another, and to the overall tasks/goals.

Third, one needs to understand the circumstances (including, but not limited to, the tasks, behaviors and intentions mentioned above) which lead to the uses of particular texts, kinds of texts, and parts of texts, and to the choice of particular media in which the texts are embodied. Understanding means, once again:

- identifying the dimensions of the situation which are salient to such choice;
- classifying these dimensions according to their relationships to one another, and to their significance for the scholars' behaviors; and,

- establishing relationships (hopefully, causal) between circumstances and behaviors.

Finally, one needs to understand how the texts are actually used by the scholars. By understanding, here we mean:

- identifying the kinds of texts and information resources which are salient to people in particular behaviors, or with particular goals, or in specific circumstances;
- identifying the characteristics of texts which are salient to people, as above;
- identifying the parts (or kinds of parts) of texts that are used by people in particular behaviors; and,
- relating the parts, kinds and media of texts that are used by people to one another, and to the goals which lead to the interaction.

The kinds of understanding of these issues that we have detailed, can then lead to:

- identification of the kinds of texts relevant to particular uses and users;
- specification of the aspects of texts relevant in particular circumstances, and the aspects of texts that are significant in supporting information seeking; and, strategies and techniques for supporting effective interaction with texts (or effective information-seeking behaviors).

Such results, then, can serve as a specification for the design of effective textual resources, of the type we have outlined above.

Our project attempts to achieve the kinds of understanding described above by observing the working behaviors of scholars in the humanities, in particular as they interact with texts and other information resources (e.g., people) in the accomplishment of their working goals. We elicit from these people, via interview, focus group, and observation, descriptions of prototypical goals and tasks in their environment, and of prototypical uses of texts. We also observe the scholars in (or ask them to report on) their behaviors as they accomplish their information and text-related tasks, and elicit from them, by interview, protocol analysis and thinking-aloud techniques, the intentions which lead them to undertake these behaviors, their judgments of their success in achieving those intentions, and in their overall tasks, and the reasons for their judgments. During observation, we pay special attention to the specific texts that are interacted with, to the parts of the texts to which the scholars specifically address themselves, to the linguistic and other textual structures which are relevant to seeking and using text, and to the intentions which underlie these interactions with texts. We have chosen scholars and environments in which texts and other information resources are available in both electronic and paper (or other media, such as photographs) forms, in order to get as wide a range of behaviors as possible, and to attempt to control for the effects of the access mechanisms and other characteristics associated with any one form of text.

Our investigation has the following structure. We begin with interviews with expert informants in the topic area(s) chosen, to get a first-order description of the four general issues of interest to us. The details of this interview are discussed in the next section.

The primary functions of this interview are:

- to gather initial data on the description and typology of tasks and goals, on the range of text interaction tasks and behaviors, and on the nature and relationships of texts to goals;
- to provide a basis for the design of the subsequent stages of the empirical investigation.

We follow with naturalistic (i.e., *in situ*) observation of scholars engaged in tasks leading to interaction with texts and other information resources. The interaction is with whatever forms and media of text the scholar finds relevant, but the opportunity to use electronic texts is explicit, and easy to realize. The form of the observation is one or more of the following:

- non-participant observation of a scholar when engaged in a relevant task;
- critical incident interviews with a scholar, recalling text-related behaviors associated with specific tasks performed in the recent past;
- concurrent self-reports by a scholar, on text-related behaviors. These may take the form of written or audio-tape diaries, or of structured reports;
- monitoring of persons engaged in interaction with electronic texts/resources.

The observation is in all cases either supplemented by, or incorporates, instruments to elicit goals, intentions, reasons and success of/for the observed behaviors. Usually, this will be in the form of interviews, or sometimes of focus groups. The data thus obtained will give us descriptions about the nature and range of the behaviors, goals, and decision processes of our sample population. They also provide a basis for the classificatory activities required by our goals. And, they help us to structure the next stage of our investigation, by providing direct information about the kinds of texts relevant to scholars in particular circumstances and the range of behaviors that we can expect; and, by suggesting hypotheses about the relationships among tasks, goals, behaviors and texts.

The last stage of empirical investigation is the controlled observation of scholars engaged in interaction with text. At this stage, we ask scholars to perform some task or tasks that we have decided, based on our earlier investigations, will lead to text-related behaviors of interest to us, in circumstances which allow us to observe and monitor their behaviors, and to some extent control the nature of the texts to which they have recourse. This stage of the investigation is of two general types: controlled and experimental. The former is based upon either a few given, or known tasks, but does not otherwise constrain the situation beyond its being in one, or a few, specific locations. The latter is based upon given tasks, and constrains the resources and behavior opportunities available to the subjects in ways that are predicted to be of interest, given the results of the previous studies. The goals of these investigations are to give us more detailed information about the specific kinds of interactions that scholars engage in, and the specific parts of texts that they use in these interactions; and, to suggest predictive, causal or other relationships

among the variables of interest to us. This set of investigations include being able to set experimental conditions in controlled environments. Apart from task performance data, monitoring data, and general observational data, we also gather data on goals, intentions, reasons for and success of interactions.

In all of these studies, we choose humanities scholars as our initial population for study. There are several reasons for this choice. Perhaps the most important is that the relationships of humanities scholars to texts, and their working habits and goals in this respect, seem to be quite different from the population of information users which has been most studied, natural scientists. In particular, texts for humanists are often the object of study, as well as sources of information about the object, and humanists typically interact with complete texts, rather than surrogates. By studying humanists, we assume that we will observe behaviors which may not have been observed in studies of scientists, and we will therefore be able to extend our general knowledge of human interactions with text substantially. Another reason for selecting scholars in the humanities is that texts in electronic form have been available to them, and used by them, for longer than for any other comparable group of scholars. The familiarity they have thus gained allows us to investigate informed choice of text medium for interaction, a key component of our study.

To date, we have completed only the first stage of our overall investigative scheme; the interviews with informants, and analysis of these data. The results of this stage include the development of a preliminary scheme for the classification of interactions of scholars with texts, and a preliminary analysis of the relationships between scholars' tasks and goals, and their interactions with texts. These results, whose primary role is to structure our further investigations, seem to us also to hold intrinsic interest, and so we present them in this paper.

### 3 Interview Methods

Our sample of scholars consists of eleven senior-level faculty members at two campuses of Rutgers University, in the Departments of English, History, and Philosophy. Ten are male, one female. All but one were interviewed in their own offices (the exception was interviewed in the office of one of the interviewers), by appointment, each interview lasting about one hour. Each entire interview was audio-taped, for subsequent transcription and analysis, and the interviewers also took notes, to supplement the transcripts, in the course of the interviews.

Each interview was structured around three general topic areas:

- professional activities and tasks the scholar is engaged in;
- goals or purposes of these tasks and activities; and,
- information-seeking and text-interaction behaviors associated with these goals and activities, with particular attention to the nature of interaction with text.

At the beginning of each interview, each respondent was asked to describe her or his working environment, in terms of typical activities. The interviews began with the

general question: "Can you tell me about the different things you do in your professional life; in other words, what are the professional activities or tasks you are engaged in?" Respondents were encouraged to talk about a variety of activities, beyond the obvious ones of teaching and research. With respect to teaching and research, an effort was made to specify further the subtasks involved in these broader activities. For example, we asked people to talk about "some of the things you do as part of your teaching(research)."

People in our study talked at length about their research goals and activities. In order to encourage more detailed descriptions of these behaviors, we asked each person to talk about a recent research project, stage by stage. For each of the tasks or activities mentioned by the scholars, we asked about the goal or purpose associated with it, and the information resources and other texts typically used. For example, one scholar discussed the task of writing a book review. We next asked him, "What do you need to do when you write a book review?" He then described the related tasks of research that go into writing the book review, with the goal of being able to evaluate the quality and thoroughness of the work. For instance, in order to achieve this goal, certain bibliographic journals are consulted. In most cases, the interviews moved through this sort of progression from discussion of activities or tasks, to associated goals and information behaviors.

In order to more precisely understand the nature of interactions with texts, we asked our respondents to reconstruct their use of a specific text, step by step. In some cases, we selected a journal from the scholar's office and asked her/him to describe what s/he does with it when it first comes in. This often led to a discussion of how the scholar evaluates the potential usefulness of specific texts within the journal, in the process of deciding what s/he will read or not read. In other cases, where the text was the object of analysis, we asked the scholar to describe in detail the ways in which the text was used, how it was analyzed, what parts were selected for analysis, how and why these parts were selected, and so on.

## 4 Results from Analysis of the Interviews

All of the interviews were transcribed from the audio-tape recordings, with the help of notes taken during the interviews. As a first step in our use of these data, a content analysis of the transcripts was conducted, in which all mentions of tasks or activities; goals, and information-seeking behaviors were identified and coded. From these data, we developed a classification scheme for the three general facets of **task**, **goal**, and **information-seeking behavior**. The general structure of these three facets is displayed in figures 1-3.

1. Teaching
2. Research
3. Project Development
4. Writing
5. Construction of Texts
6. Service

Figure 1. Tasks of Humanities Scholars

1. Maintaining Personal Knowledge Structures
2. Identifying
3. Learning
4. Discovering
5. Judging
6. Evaluating
7. Finding
8. Understanding/Contextualizing
9. Creating/Maintaining Professional Identity

**Figure 2. Goals of Humanities Scholars**

At this point, we are beginning to use the classification scheme in our analysis of how humanities scholars use texts. The data suggest that these scholars interact with multiple aspects of texts, in support of multiple and overlapping goals.

1. Mode of Behavior
  - 1.1 Interpersonal Communication
  - 1.2 Interaction with Texts
    - 1.2.1 Parts of Texts
    - 1.2.2 Kinds of Texts
2. Method of Interaction
3. Orientation to Texts

**Figure 3. Information-Seeking Behaviors of Humanities Scholars**

In Cool, et al. (1993), we discussed one particular type of interaction with text - that with the goal of *evaluating* usefulness. In order to keep up to date in their field, and to identify new problem areas, the humanities scholars we interviewed typically scan a wide periodical literature. In order to select useful articles to read from this wide body of literature, many of the scholars look first at characteristics of the text such as title or author, to judge the topical relevance of a specific text. These document characteristics are most useful in identifying texts which are *not* interesting to the scholar. If the text appears interesting or relevant based on title, then other criteria may be employed in the evaluation process. The title of an article may appear to be on a topic of interest to the scholar, but the author may be unknown; or the author may be a well-known scholar, but the specific topic of the article may not be readily identifiable. In these ambiguous cases, the scholars frequently examine an additional attribute of the text, such as the footnotes, or the acknowledgements, to identify the author's specific point of view, or reference group. From this aspect of the text, the scholar can tell something about the way in which the topic is treated in the text, and about who the author is. In this example behavior, the evaluation of the text involves more than an assessment of topical relevance; it involves

establishing the author's "point of view", determining the credibility or authority of the author, and determining the scholarly community the text is intended to address.

We can also give an example of another type of interaction, that with the goal of *discovering*, to illustrate how scholars in our study use texts and parts of texts which are themselves objects of analysis. For example, one of our participants is a medievalist who is interested in discovering examples of empathy in 12th and 13th century texts. Discovering these examples is difficult because empathy is an abstract concept, and the word itself did not exist at the time the texts were written. This scholar came upon the idea of studying empathy after he found it portrayed in a picture of Francis of Assisi receiving the stigmata. As he says, "this is a vivid portrayal of empathy in action because Francis is literally empathizing with Christ. And these very words are used by people writing not long after his death to describe him." After identifying empathy as a topic, the early stages of his research involve scanning whole texts widely, "everything is sight about Francis", looking for places in the text where empathy might be revealed. He is not looking for the word empathy, but for parts of texts such as episodes, or passages, which may illustrate it. As he says:

And whether you know it or not, you have in mind a list of things to watch out for. Words, relationships, episodes. For example, Francis preaching to the birds. Or Francis domesticating the wolf. You want to know more about those. These are very famous episodes. Why did he do it? Why do the texts say he did it? You look at the language in the text. He preached to the birds. Does this show that he really felt God in God's creatures? That he empathized with all of nature?

This scholar is particularly interested in discovering the limits of empathy, or with whom empathetic participation is possible. In order to discover this, he moves from the analysis of specific parts of texts back to the text as a whole. Here he looks across the entire text for the word "love" in various forms, in order to identify between whom loving relationships are possible.

You find love. Different kinds of love. You find various people mentioned with whom loving relationships do not exist. You have to look at different relationships to figure out with whom loving relationships can exist. What words are used by Augustine to describe his relations with concubines? What words are used to describe his relationship with his mother? . . . his relations with his nurses when he was an infant? Then, what words does he use to describe their dealings with him? Does he use the same words? The answer is no.

These two examples indicate two very different requirements on text representation and structuring for supporting effective interaction with text. The former shows that formal characteristics of secondary texts, normally used for such purposes as identification and descriptive cataloging, can be, and are routinely used for highly conceptual purposes. Thus, such characteristics will need to be represented and related to one another, and other characteristics of texts, in forms that can be used in ways other than for bibliographic specification, or as retrieval keys.

The latter example suggests that scholars dynamically and iteratively construct instantiations of tropes, through their interactions with texts, which they progressively search for and modify. This implies that we require facilities for dynamic, user-defined, large-scale patterns in text. Furthermore, such patterns, as created, should be usable for partial, rather than exact matching on the text collection as a whole, and should be capable of being related to other portions of the texts through dynamically established clusters or links. These requirements have strong implications both for what we consider as atomic elements in a text, and for levels of agglomeration of those elements. Just word representation and word search is clearly inadequate to these purposes, as are pre-defined patterns or structures.

Both of these examples also show explicitly the interactive nature of information-seeking behaviors, and that such interaction is deeply dependent upon the knowledge and goals of the human actor.

## 5 Conclusions

Since this is very much a report on work in progress, here we discuss how the results to date respond to the purposes of our study as a whole, and what the next stages will entail. Our preliminary results have responded to some extent to all four of our original goals. That is, we have some understanding of the range of scholars' goals and tasks, and a preliminary classification of these tasks; we have identified and begun to classify some interactive behaviors with texts; we have some very preliminary understanding of some of the dimensions of the general scholarly situation which lead to the use of particular texts and modes; and, we have begun to identify various characteristics of texts in terms of their interactive uses. In particular, we now have some ideas about some characteristics of texts which will need to be represented in electronic resources in order to support the kinds of interactions that we have observed, and also some ideas about the nature of the access and other tools which a system would need to support effective interaction between scholars and texts.

Thus, we can say that, although our results thus far are not terribly startling in terms of their implications, they nevertheless support the assumptions of our general program. That is, that the design of electronic textual resources should be based on, and embedded in, the tasks and information-seeking behaviors of those who will be using the texts. Our next series of studies, which are now underway, will build upon our results to date, and extend them in specific ways. Specifically, the following studies are either now beginning, or will be carried out in the near term.

First, we are engaging in more interviews, with more scholars in the field, and are doing fuller analyses of these data. This responds to the clear problem that we may not have a sufficiently representative sample of tasks, goals, behaviors and uses in our preliminary data. Second, we have begun equivalent studies of other user groups, including academic natural scientists, and students, both in engineering and in the humanities. These studies are explicitly intended to discover whether there are common classes of tasks, goals, and interactive behaviors among the varieties of users, and also to identify the conditions under which significant differences arise. We are also beginning studies in which we observe people in their interactions with texts of various sorts, rather than interviewing them about such interactions. These studies will allow us to

investigate more fully when and why one form of text is preferred to another, which in turn will lead to more principled connections between goals and behaviors than we have been able to establish to date. Finally, we are beginning to design studies in which we set specific kinds of tasks to be performed in a controlled environment, in order both to observe sequences of behaviors, and to test predictions of behaviors based upon our earlier analyses.

We plan then to build a prototype electronic textual resource, based upon the results obtained from these studies, which will be made available to a variety of users, and which will serve as a testbed for continued cycles of formative evaluation and design. This we hope to do this within the context of the Center for Electronic Texts in the Humanities, allowing us to provide a new resource for the community of scholars in the humanities, and to respond to those scholars' uses of texts in a principled and generalizable manner.

## References

- [1] Bates, M.J. "Humanities scholars' use of online searching in their scholarship", in *ASIS '93. Proceedings of the 56th ASIS Annual Meeting*, Learned Information, Inc., Medford NJ, 1993, 283.
- [2] Chu, C.M. (moderator) "How humanists think: theoretical and cognitive aspects", in *ASIS '93. Proceedings of the 56th ASIS Annual Meeting*, Learned Information, Inc., Medford NJ, 1993, 282.
- [3] Cool, C., N.J. Belkin, O. Frieder, P. Kantor, "Characteristics of text affecting relevance judgments", in M.E. Williams, (ed.), *Proceedings of the Fourteenth National Online Meeting*, Learned Information, Inc., Medford NJ, 1993, 77-84.
- [4] *Humanists at work: disciplinary perspectives and personal reflections*. University Library, University of Illinois at Chicago, 1989.
- [5] Wiberley, S.E. "Revisiting humanists at work", in *ASIS '93. Proceedings of the 56th ASIS Annual Meeting*, Learned Information, Inc., Medford NJ, 1993, 282.
- [6] Walker, D.E. "The organization and use of information: contributions of information science, computational linguistics, and artificial intelligence", *Journal of the American Society for Information Science*, 32 (5), 1981, 347-363.

**SECTION 4**

**TOPICS, METHODS AND FORMALISMS IN SYNTAX,  
SEMANTICS AND PRAGMATICS**

# Some Recent Trends In Natural Language Processing\*

Aravind K. Joshi

Department of Computer and Information Science  
University of Pennsylvania, Philadelphia, PA 19104  
*e-mail:* [joshi@linc.cis.upenn.edu](mailto:joshi@linc.cis.upenn.edu)

## Abstract

In this paper, I will describe a few aspects of Natural Language Processing (NLP), specifically some recent trends. These topics concern some new avenues of research in grammars and parsing, and statistical approaches to NLP, a relatively new trend in NLP.

## 1 Introduction

Language (spoken and written) is central to all aspects of our communication. Therefore natural language processing systems (NLP), both current and future, are bound to play a crucial role in our communication with machines and even among ourselves. NLP systems include systems for speech recognition, language understanding and language generation. Spoken language systems are those that integrate speech and language systems. Such systems will provide and to some extent already do so, an interface to databases and knowledge bases; for example, an airline information and reservation system, expert systems for scheduling, planning, and maintenance, among others. Text processing and message understanding systems are useful for extracting information from texts and formatting it in a variety of ways for further use. Language communication often occurs in two or more languages. Multilingual NLP has applications to a variety of multilingual tasks such as providing aids for translating foreign language correspondence, translating equipment manuals, and speech-to-speech translation in limited domains, among others. Finally, natural language in conjunction with graphic/animation modality provides very useful cooperative interfaces, especially in the instructional domains.

Many major topics have been omitted, all of which are very important to NLP. I have not discussed speech recognition and synthesis at all, and in the language area, I have not discussed many aspects of discourse structure, which are crucial to natural language understanding and generation and their applications to cooperative interfaces. I have also not discussed the very important area of machine translation.

---

\*This work was partially supported by NSF Grant DCR-84-10413, ARO Grant DAAL03-87-0031, and ARPA Grant N0014-85-K0018. This is a slightly revised version of an invited paper presented at the International Conference on Pattern Recognition and Digital Technologies held at the Indian Statistical Institute, Calcutta, India, December 1993.

## 2 Grammars and Parsers

Language has hierarchical structure at various levels, in particular at the sentence level, which is the level we will be concerned with in this section. Almost every NLP system has a grammar and an associated parser. A grammar is a finite specification of a potentially infinite number of sentences, and a parser for the grammar is an algorithm that analyzes a sentence and assigns one or more structural descriptions to the sentence according to the grammar, if the sentence can be characterized by the grammar. A structural description is a record of the derivational history of the sentence according to the grammar. The structural descriptions are necessary for further processing, for example, for semantic interpretation. Chomsky's work on formal grammars in the late 50s was the beginning of the investigation of mathematical and computational modeling of grammars (1959). He introduced a hierarchy of grammars (finite state grammars, context-free grammars, context-sensitive grammars, and unrestricted rewriting systems) and investigated their linguistic adequacy.

Many NLP systems are based on context-free grammars (CFG). We will briefly describe CFGs. A CFG,  $G$ , consists of a finite set of non-terminals (for example,  $S$ : sentence;  $NP$ : noun phrase;  $VP$ : verb phrase;  $V$ : verb;  $ADV$ : adverb), a finite set of terminals (for example, *Harry*, *peanuts*, *likes*, *passionately*), and a finite set of rewrite rules of the form  $A \rightarrow W$ , where  $A$  is a non-terminal and  $W$  is a string of zero or more non-terminals and terminals.  $S$  is a special non-terminal called the start symbol. A derivation in a grammar begins with  $S$ , the start symbol.  $S$  is rewritten as a string of non-terminals and terminals, using a rewrite rule applicable to  $S$ . The new non-terminals are then rewritten according to the rewrite rules applicable to them, until no further rules can be applied. It is easy to see that the sentence *Harry likes peanuts passionately* can be generated by the grammar. A derivation starts with the start symbol  $S$ . This symbol is then rewritten as the string  $NP\ VP$ . These two symbols are now rewritten (in any order) as the strings *Harry* and *VP ADV* respectively. The symbol *VP* is rewritten as the string *V NP* and *ADV* is rewritten as *passionately*. Finally, *V* is rewritten as *likes* and *NP* is rewritten as *peanuts*. This derivation (derivation tree) corresponds to the sentence *Harry likes peanuts passionately*.

A finite-state grammar is like a CFG, except that the rewrite rules are of the form  $A \rightarrow aB$  or  $A \rightarrow a$ , where  $A$  and  $B$  are non-terminals and  $a$  is a terminal symbol. Finite-state grammars have been shown to be inadequate for modeling natural language structure. This is because there are dependencies that hold at unbounded distances. A context-sensitive grammar is also like a CFG, except that the rewriting of a non-terminal is dependent on the context surrounding the non-terminal, unlike the rewrite rules in CFG where the rewriting is context-independent. Context-sensitive grammars appear to be adequate for describing natural language structures. However, the entire class of context-sensitive grammars appears to be too powerful in the sense that it is not constrained enough to characterize just the structures that arise in natural language.

CFGs, as defined above, are inadequate for a variety of reasons and need to be augmented. The two main reasons are as follows: (i) The information associated with a phrase (a string of terminals) is not just the atomic symbols used as non-terminals. A complex bundle of information (sets of attribute-value pairs, called feature structures) has to be associated with strings, the syntactic category of the phrase being only one

such feature, for example. Appropriate structures and operations for combining them are needed together with a CFG skeleton. (ii) The string combining operation in a CFG is concatenation; that is, if  $u$  and  $v$  are strings,  $v$  concatenated with  $u$  gives the string  $w = uv$ , that is,  $u$  followed by  $v$ . More complex string combining as well as tree combining operations are needed to describe various linguistic phenomena. We will illustrate these two kinds of augmentations by some simple examples.

## 2.1 Mildly Context-Sensitive Grammars

In any mathematical or computational grammar, a wide range of dependencies among the different elements in the grammar have to be described. Some examples of these dependencies are as follows:

1. Agreement features such as person, number, and gender. For example, in English, the verb agrees with the subject in person and number.
2. Verb subcategorization, in which each verb specifies one (or more) subcategorization frames for their complements. For instance, *sleep* does not require any complement (as in *Harry sleeps*), *like* requires one complement (as in *Harry likes peanuts*), *give* requires two complements (as in *Harry gives Susan a flower*), and so forth.
3. Sometimes the dependent elements do not appear in their normal positions. In

*Who<sub>i</sub>; did John invite e<sub>i</sub>*

where  $e_i$  is a stand-in for  $who_i$ ,  $who_i$  is the filler for the gap  $e_i$ . The filler and the gap need not be at a fixed distance. Thus in *who<sub>i</sub>; did Bill ask John to invite e<sub>i</sub>*, the filler and the gap are more distant than in the previous sentence.

4. Sometimes the dependencies are nested. In German, for example, one could have

*Hans<sub>i</sub>; Peter<sub>j</sub>; Marie<sub>k</sub>; schwimmen<sub>k</sub>; lassen<sub>j</sub>; sah<sub>i</sub>*  
(Hans saw Peter make Marie swim)

where the nouns (arguments) and verbs are in nested order, as the subscripts indicate.

5. However, in Dutch, these dependencies are crossed, as for example, in

*Jan; Piet; Marie<sub>k</sub>; zag; laten; zwemmen<sub>k</sub>*  
(Jan saw Piet make Marie swim).

There are, of course, situations where the dependencies have more complex patterns. Precise statements of such dependencies and the domains over which they operate constitute the major activity in the specification of a grammar. Mathematical and computational

modeling of these dependencies is one of the key areas in natural language processing. Many of these dependencies (for example, the crossed dependencies discussed above) cannot be described by context-free grammars. This is easily seen from the well-known fact that CFGs are equivalent to the so-called push-down automata (PDAs) which have the storage discipline—last in first out. PDAs therefore can characterize nested dependencies but not the crossed dependencies.

In the context-free grammar (CFG), the dependency between a verb (*likes*) and its two arguments [subject (NP) and object (NP)] is specified by means of two rules of the grammar. It is not possible to specify this dependency in a single rule without giving up the VP (verb phrase) node in the structure. That is, if we introduce a rule,  $S \rightarrow NP\ V\ NP$ , then we can express the dependency in one rule, but then we cannot have VP in our grammar. Hence, if we regard each rule of a CFG as specifying the domain of locality, then the domain of locality for a CFG cannot locally (that is, in one rule) encode the dependency between a verb and its arguments, and still keep the VP node in the grammar.

We will now describe briefly one grammar formalism whose domain of locality is larger than that of a CFG. In the tree-adjoining grammar (TAG), each word is associated with a structure (tree) (the word serves as an *anchor* for the tree) which encodes the dependencies between this word and its arguments (and therefore indirectly its dependency on other words which are anchors for structures that will fill up the slots of the arguments). Thus for *likes*, the associated tree encodes the arguments of *likes* (that is, the two NP nodes in the tree for *likes*) and also provides slots in the structure where they would fit. The trees for *Harry* and *peanuts* can be substituted respectively in the subject and object slots of the tree for *likes*. The tree for *passionately* can be inserted (adjoined) into the tree for *likes* at the VP node. The derivation in a TAG grammar is quite different from the derivation in a CFG. The derivation tree will be a record of the history of the various adjoining and substitutions carried to produce the derived tree. In a TAG, the entire grammar consists of lexical items and their associated structures. There are universal operations, substitution and adjoining which describe how structures can be combined (Joshi 1985, 1987; Joshi and Schabes 1991).

## 2.2 Parsing Complexity

A parser for a grammar is an algorithm that assigns to a sentence one or more structural descriptions according to the grammar, if the sentence is generable by the grammar. Parsing of sentences according to different grammars and the complexity of this process are important research areas in NLP. For a CFG a number of parsing algorithms are known and the time required to parse a sentence of length  $n$  is at most  $Kn^3$  where  $K$  depends on the size of the grammar. This result extends to almost all CFG-based grammars used in NLP. The constant  $K$  can become very large however. In practice, of course, the worst case complexity is really not the important measure. Most parsers perform much better than the worst case on typical sentences. There are no mathematical results, as yet, to characterize the behavior on typical sentences. Grammars that are more powerful than CFG are, of course, harder to parse, as far as the worst case is concerned. The grammars in the class of Mildly Context-Sensitive Grammars discussed earlier can all be parsed in polynomial time just as CFG; however, the exponent for  $n$  is 6 instead of 3.

A crucial problem in parsing is not just to get all possible parses for a sentence but to rank the parses according to some criteria. If a grammar is combined with statistical information, then that information can be used to provide this ranking. This is exactly what is done in many spoken language systems, that is, systems that integrate speech recognition and language processing.

In our discussion so far, we have been assuming that the parser only handles complete sentences and the parser either succeeds in finding the parse(s) for a sentence or it fails. In practice, we want the parser to be flexible—that is, it should be able to handle fragments of sentences—and it should fail gracefully—that is, it should provide as much analysis as possible for as many fragments of the sentence as possible, even if it cannot glue all the pieces together. A parser with such properties based on the idea of deterministic parsing has been described in Marcus (1980) and used in the construction of a large corpus of parsed text, a tree bank.

Finally, the actual grammars in major NLP systems are large, but even with this large size their coverage is not adequate. Building the grammar by hand soon reaches its limit and there is no guarantee that it will be increasingly better in coping with free text (say, text from a newspaper) by continuing to build it manually. Increasing attention is being paid now to automatically acquiring grammars from a large corpus (Brill et al. 1990).

### 3 Statistical Approaches to Natural Language Processing

There is a long history of modeling language statistically. After all, some words occur more frequently than other words (for example, *the* occurs more frequently than *man*, which occurs more frequently than *aardvark*) some two-word sequences appear more frequently than some other two-word sequences (for example, *a man* occurs more frequently than *old man*, which occurs more frequently than *green man*), and so forth. Hence, it is reasonable to believe that language can be modeled statistically. A specific proposal along these lines was made by Shannon (1948). He viewed the generation process as modeled by stochastic processes, in particular, a Markov process. For our present purpose, we will characterize sentence generation by a finite state machine (Fig. 1). Given a state diagram, we generate a sentence by starting with the initial state and then traversing the diagram from state to state and emitting the word labeling the arc between a pair of states. The process ends when we reach the final state. A probability is assigned to each state transition together with the emitted symbol, that is, to a triple  $(S_i, a_j, S_k)$  representing the transition from state  $S_i$  to state  $S_k$  emitting the symbol  $a_j$ . Although such machines are clearly relevant to modeling language statistically, Chomsky (1957) rejected the finite state machine characterization as inappropriate for modeling grammars, for the following reason: In Fig. 1, *lives* is four words away from *man*, if we did not follow the loop at  $S_4$ . Hence the dependency between these two words can be captured by the state sequence from  $S_2$  to  $S_6$ . However, in the sentence *The man who the woman Harry met yesterday telephoned lives in Philadelphia* (one that is a bit difficult to process but grammatical, and not generable by the machine in Fig. 1), *lives* is now seven words away from *man*. Since more clauses can be embedded and each clause can be lengthened by adding adjectives or adverbs, the distance between *lives* and *man* can be made arbitrarily large and thus the number of states required to model language

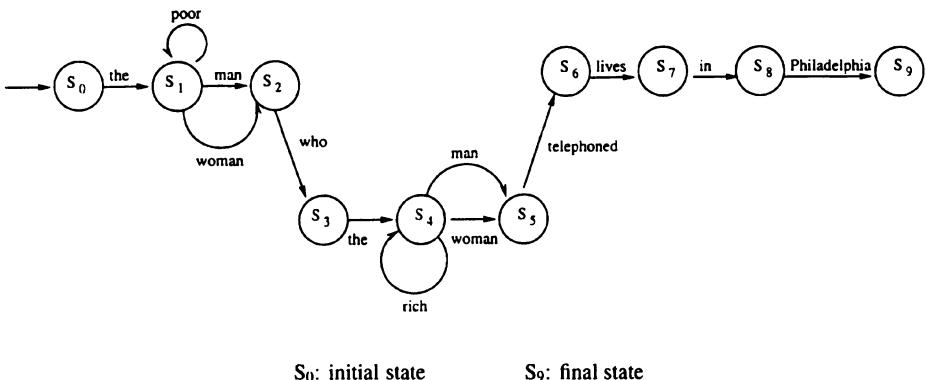


Figure 1: A finite state machine generating sentences

cannot be bounded. Hence a finite state machine is inadequate. Chomsky also rejected the possibility of associating the probability of a sentence with its grammaticality (the higher the probability, the higher the grammaticality of the sentence). This is because if we order the sequences of a given length (there will be  $W^n$  such sequences, if  $W$  is the number of words and  $n$  is the length of the sequences) according to the probabilities of the sequences then it will not be possible to sort out grammatical and ungrammatical sequences on the basis of this ranking (Chomsky 1957). He then developed structural models, such as the phrase structure grammar and transformational grammar, which formed the basis for almost all of the work in mathematical and computational linguistics up until the present.

Although Chomsky rejected the statistical models, he commented

Given the grammar of language, one can study the use of the language statistically in various ways; and the development of probabilistic models for the use of language (as distinct from the syntactic structure of language) can be rewarding... One might seek to develop a more elaborate relation between statistical and syntactic structure than the simple order of approximation model we have rejected. I would certainly not care to argue that any such relation is unthinkable, but I know of no suggestion to this effect that does not have obvious flaws..." (Chomsky 1957)

Harris (1957) proposed a transformational theory motivated by the considerations of normalizing sentence structures (for the purpose of discourse analysis) so that the relevant co-occurrences among words can be stated in a local manner. Very roughly speaking, under this view, *The man who Harry met yesterday lives in Philadelphia* is made up of S1: *The man lives in Philadelphia* and S2: *who Harry met* (which is a transformed version of S3: *Harry met the man*, with S1 and S3 sharing *the man*) and so on. There are clearly ‘meaningful’ statistical dependencies between *lives* and the subject

noun *man* and the object of *in*, namely, *Philadelphia*, and between *met* and *Harry*, the subject of *met*, and *man* the object of *met*, but not ‘meaningful’ statistical dependencies between *lives* and *yesterday* or *met yesterday* (the one-word and two-word sequences before *lives*) and so on.

Although statistical approaches did not play a significant role in mathematical or computational linguistics, it is clear that the idea of somehow combining structural and statistical information was already suggested as early as the late 50s. Now in the 90s, we see a resurgence of these early ideas. There are two key reasons for this renewed interest. First we now have some formal frameworks which appear to be suitable for combining structural and statistical information in a principled manner and second, there is now the possibility of using very large corpora, annotated in various ways that can be used for reliably estimating the various statistics needed to deduce linguistic structure (see Brill et al. 1990).

Hidden Markov Models (HMM) have played a crucial role in speech recognition. HMMs are derived from the theory of probabilistic functions of finite state Markov chains (Baum and Petrie 1966; Paul 1990). HMM’s were introduced in the speech recognition domain in the early 80s and became very popular in the late 80’s. They have also found use in the spoken language systems, i.e., systems that integrate speech and natural language. As we have already pointed out finite state models are not adequate for modeling the structure of natural language, more powerful models such as context-free grammars and beyond are needed. The parameter estimation techniques for HMMs have been extended to these more powerful models also (Pereira and Schabes 1992; Lari and Young 1990; Jelinek et al. 1990).

HMMs are equivalent to finite state (stochastic) grammars (regular grammars). Finite state grammars are not adequate to model certain aspects of language, in particular the recursive aspects, as described earlier. Hence, it is useful to consider more powerful grammars such as context-free grammars, i.e., consider stochastic context-free grammars. The forward-backward algorithm for training HMMs can be extended to stochastic context-free grammars (Jelinek et al. 1990; Pereira and Schabes 1992; Lari and Young 1990). In this case, it is often referred to as the inside-outside algorithm. We assume that the context-free grammar is in the Chomsky normal form, i.e., the rules of the grammar are of the form

$$A \rightarrow BC$$

$$A \rightarrow a$$

where  $A, B$ , and  $C$  are non-terminals in the grammar and  $c$  is a terminal symbol. Let  $w = a_1, a_2, \dots, a_n$  be the string of words (observation sequence). Training this model consists of determining a set of grammar rules given a training set of sentences (strings of words),  $w_1, w_2, \dots, w_n$ . Instead of computing the forward and backward probabilities as in the case of HMMs, we compute inside and outside probabilities. Very roughly the inside probability is a computation that proceeds from bottom to top in the derivation tree and the outside probability computation proceeds from top down in the derivation of a string. For a simple description of this algorithm and its use in the reestimation of the parameters (the probabilities associated with the rules), see Jelinek et al. (1990), Pereira and Schabes (1992), and Lari and Young (1990). A similar inside-outside algorithm for reestimation has been designed and implemented for the tree adjoining grammars (Schabes 1991).

We will now give a few examples to show how structural and statistical information can be integrated. Context-free grammars (CFG) have been used extensively in modeling grammars. Each rule (production) in a CFG can be associated with a probability of its use. Thus, given a CFG with rules: (R1)  $S \rightarrow NP\ VP$  (0.9), (R2)  $S \rightarrow NP\ NP\ V$  (0.1), (R3)  $VP \rightarrow V\ NP$  (0.7), (R4)  $VP \rightarrow V$  (0.3), we have associated probabilities with each of the rules. The probabilities of all rules associated with a given non-terminal add up to 1. The probability of a sentence (more precisely the derivation of the sentence in the grammar) is simply the product of the probabilities of each rule in the derivation because the grammar is CFG and the application of a rule depends only on the non-terminal on the left hand side of a rule and not on the context in which this non-terminal appears in a derivation. Probabilistic parsing methods and methods for estimating the probabilities of the rules from a training corpus are given (see references above). . By making the probability associated with each rule somewhat context-dependent, for example, making it dependent on the preceding rule in the derivation, considerable improvement in the estimation of the probabilities and performance of the parser (in terms of getting correct parsers) can be achieved (Magerman and Marcus 1991).

As we have seen earlier, the really ‘meaningful’ statistical dependencies are between words (lexical items) mediated most likely by grammatical relations. For example, there will be ‘meaningful’ statistical dependencies between the verb *eats*, and the lexical items that can appear as subject and object of *eats*. CFGs and their generalizations are not directly based on lexical items, that is, they are not lexicalized, and in general, cannot be lexicalized (Joshi and Schabes 1991). Lexicalized grammars, as described earlier, are more appropriate for integrating structural and statistical information in a uniform manner.

Two dependent words in a sentence can be at an arbitrary distance apart, as we have seen earlier. Hence, this dependency cannot be captured by one-word, two-word, three-word and  $n$ -word frequencies, for some fixed  $n$  (that is, uni-gram, bi-gram, tri-gram and  $n$ -gram statistics). However, in many situations these statistics work surprisingly well in determining some aspects of language structure. Tri-gram frequencies (of parts of speech—that is, syntactic categories—and not words directly) have been used very successfully for discovering an optimum assignment of parts of speech to words (Church 1988, DeRose 1988). Almost all words are lexically ambiguous, that is, they belong to more than one category. For example, *table* is either a noun (N) or a verb (V); *pale* is either an adjective (ADJ) or an adverb (ADV); *see* can be a verb (V), an interjection (UM), or a noun (with capital S); *round* can be an adjective (ADJ), noun (N), verb (V), or an adverb (ADV), and so forth. Church’s program uses a linear time dynamic programming algorithm to find an assignment of parts of speech optimizing the product of: (i) probability of observing a part of speech  $i$ , given the word  $j$ , and (ii) probability of observing part of speech  $i$ , given two previous parts of speech. Probability estimates are obtained by training on a tagged corpus (such as the well-known Tagged Brown Corpus; Francis and Kučera 1982). Error rates of only 3% to 4% have been reported [8], which compare very well with the error-rate of human annotators. Similar techniques have been used to locate simple noun phrases with high accuracy.

Statistical techniques in conjunction with large corpora (raw texts or annotated in various ways) have also been used to automatically acquire other linguistic information such as morphological information (that is, parts of words such as prefixes and suffixes

and inflected forms), subcategorization information (see the earlier section on grammars and parsers for subcategorization information), semantic classes (such as classification of nouns, based on what predicates they go with; compound nouns such as *jet engines*, *stock market prices*; classification of verbs, for example, *to know* describes a state of the world, while *to look* describes events and so on), and, of course, grammatical structure itself as we have already mentioned (Magerman and Marcus 1991, Brill 1990, Brent and Berwick 1991, Brent 1991, Hindle 1990, Smadja and McKeown, 1990). Such results have opened up a new direction of research in NLP, which is often described as corpus-based NLP.

It should be clear from the previous discussion that, for the development of corpus-based NLP, very large quantities of data are required (the Brown Corpus from the 60s is about 1 million words). Researchers estimate that about 100 million words will be required for some tasks. The technologies that will benefit from corpus-based NLP include speech recognition and synthesis, machine translation, full-text information retrieval, and message understanding, among others. The need for establishing very large text and speech databases annotated in various ways is now well understood. It is recognized that no single organization can afford to create enough linguistic data even for its own research and development, let alone for the needs of the research community at large. This need, together with the size of the database and the need for sharing it, has been the key motivation for the plans for setting up a Linguistic Data Consortium (LDC) by DARPA (Liberman 1991). Initial plans of the LDC call for the collection of raw text (naturally occurring text from a wide range of sources, 5 to 10 billion words); annotated text (syntactic and semantic labeling of some parts of raw text, upwards of 20 million words); raw speech (spontaneous speech from a variety of interactive tasks, 400 hours, 2000 speakers); read speech (1,000 hours, 10,000 speakers); annotated speech (phonetic and prosodic labeling, 20 hours); a lexicon (a computational dictionary of 200,000 entries plus a term bank containing, for example, geographical, individual, and organizational names, 200 to 300 thousand entries); and a broad coverage computational grammar. The LDC will also develop a variety of sharable tools. Some examples in the speech area are: programs for segmentation of speech, alignment of speech and text, prediction of pronunciation options from orthographic transcription. Some examples from text are: a program for breaking text into sentences, a statistical parts-of-speech tagger, an efficient program for computing  $n$ -gram statistics and a variety of other statistics over very large corpora (Liberman 1991).

## References

- [1] L.F. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37:1554–1565, 1966.
- [2] M.R. Brent. Automatic semantic classification of verbs from their syntactic contexts: An implemented classifier for stativity. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, pages 222–226. Association for Computational Linguistics, Morristown NJ, 1991.

- [3] M.R. Brent and R. Berwick. Automatic acquisition of subcategorization frames from tagged text. In *Proceedings of DARPA Workshop on Spoken Language Systems*, pages 342–345. Morgan Kauffman, Palo Alto, CA, 1991.
- [4] E. Brill. Personal communication.
- [5] E. Brill, D. Magerman, M. Marcus, and B. Santorini. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of DARPA Workshop on Spoken Language Systems*, pages 275–282. Morgan Kauffman, Palo Alto, CA, 1990.
- [6] N. Chomsky. *Syntactic Structures*. Mouton & Co., S. Gravenhage, 1957.
- [7] N. Chomsky. On certain formal properties of grammars. *Informatics and Control*, 5:137–167, 1959.
- [8] K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143. Association for Computational Linguistics, Morristown, NJ, 1988.
- [9] S. DeRose. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1), 1988.
- [10] W. Francis and H. Kučera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin Company, Boston, 1982.
- [11] Z.S. Harris. Co-occurrence and transformation in linguistic structure. *Language*, 3:283–340, 1957.
- [12] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the Association for Computational Linguistics Conference*, pages 268–275, Association for Computational Linguistics, Morristown NJ, 1990.
- [13] F. Jelinek, J.D. Lafferty, and R.L. Mercer. Basic methods of probabilistic grammars. Technical report, IBM, Yorktown Heights, NY, 1990.
- [14] A.K. Joshi. How much context-sensitivity is necessary for characterizing structural descriptions—Tree adjoining grammars. In A. Zwicky D. Dowty, L. Karttunen, eds., *Natural Language Processing: Theoretical, Computational and Psychological Perspectives*. Cambridge University Press, Cambridge, England, 1985.
- [15] A.K. Joshi. An introduction to tree adjoining grammars. In A. Manaster-Ramer, ed., *Mathematics of Language*, pages 87–114. John Benjamin, Amsterdam, 1987.
- [16] A.K. Joshi and Y. Schabes. Flexible phrase structure and coordination. In *Proceedings of DARPA Workshop on Spoken Language Systems*, pages 195–199. Morgan Kauffman, Palo Alto, CA, 1991.
- [17] K. Lari and S.J. Young. Applications of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 5:237–257, 1990.

- [18] M. Liberman. Guidelines for the linguistic data consortium. Draft Proposal for DARPA, May 1991.
- [19] D. Magerman and M. Marcus. Pearl: A probabilistic chart parser. In *Proceedings of the Fifth Conference of the European Association for Computational Linguistics*, pages 15–20. Association for Computational Linguistics, Morristown NJ, 1991.
- [20] M. Marcus. *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge MA, 1980.
- [21] D.B. Paul. Speech recognition using the hidden markov model. *The Lincoln Laboratory Journal*, 3(1):41–62, Spring 1990.
- [22] F. Pereira and Y. Schabes. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the DARPA Speech and Natural Language Systems Workshop*, Arden House, NY, 1992. Morgan Kauffman.
- [23] Y. Schabes. A inside-outside algorithm for estimating the parameters of a hidden stochastic context-free grammar based on Earley’s algorithm. In *Second Workshop on Mathematics of Language (MOL)*, Yorktown Heights, NY, May 1991.
- [24] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27(379), 1948.
- [25] F. Smadja and K. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the Association for Computational Linguistics Conference*, pages 252–259, Association for Computational Linguistics, Morristown NJ, 1990.

# Two Principles of Parse Preference\*

Jerry R. Hobbs and John Bear  
Artificial Intelligence Center  
SRI International  
*e-mail: hobbs@ai.sri.com*

## 1 Introduction

The DIALOGIC system for syntactic analysis and semantic translation has been under development for over ten years, and during that time it has been used in a number of domains in both database interface and message-processing applications. In addition, it has been tested on a number of sentences of linguistic interest. Built into the system are facilities for ranking parses according to syntactic and selectional considerations, and over the years, as various kinds of ambiguity have become apparent, heuristics have been devised for choosing the preferred parses. Our aim in this paper is first to present a compendium of many of these heuristics and second to propose two principles that seem to underlie the heuristics. The first will be useful to researchers engaged in building grammars of similarly broad coverage. The second is of psychological interest and may be a guide for estimating parse preferences for newly discovered ambiguities for which we lack the experience to decide among on a more empirical basis.

The mechanism for implementing parse preference heuristics in DIALOGIC is quite simple. Terminal nodes of a parse tree acquire a score (usually 0) from the lexical entry for the word sense. When a nonterminal node of a parse tree is constructed, it is given an initial score which is the sum of the scores of its child nodes. Various conditions are checked during the construction of the node and, as a result, a score of 20, 10, 3, -3, -10, or -20 may be added to the initial score. The score of the parse is the score of its root node. The parses of ambiguous sentences are ranked according to their scores. Although simple, this method has been very successful. In this paper, however, rather than describe the heuristics in terms this detailed, we will describe them in terms of the preferences among the alternate structures that motivated our scoring schemes.

While these heuristics have arisen primarily through our everyday experience with the system, we have done small empirical studies by hand on some of the ambiguities, using several different kinds of text, including some from the Brown corpus and some transcripts of spoken dialogue. We have counted the number of occurrences of potentially ambiguous constructions that were in accord with our claims, and the number

---

\*The authors would like to express their gratitude to Paul Martin, who is responsible for discovering some of the heuristics, and to Mark Liberman for sending us some of the data. The research was funded by the Defense Advanced Research Projects Agency under Office of Naval Research contract N00014-85-C-0013, and by a gift from the Systems Development Foundation.

of occurrences that were not. Some of the constructions were impossible to find, not only because they occur so rarely but also because many are very difficult for anyone except a dumb parser to spot. But in every case where we found examples, the numbers supported our claims. We present our findings below for those cases where we have begun to accumulate a nontrivial number of examples.

## 2 Brief Review of the Literature

Most previous work on parse preferences has concerned itself with the most notorious of the ambiguities—the attachment ambiguities of postmodifiers. Among the first linguists to address this problem was Kimball (1973). He proposed several processing principles in an attempt to account for why certain readings of ambiguous sentences were more salient than others. Two of these principles were Right Association and Closure.

In the late 1970s and early 1980s there was a great deal of work among linguists and psycholinguists (e.g., Frazier and Fodor, 1979; Wanner and Maratsos, 1978; Marcus, 1980; Church, 1980; Ford, Bresnan, and Kaplan, 1982) attempting to refine Kimball's initial analysis of syntactic bias and proposing their own principles governing attachment. Frazier and Fodor proposed the principles of Minimal Attachment and Local Association. Church proposed the A-over-A Early Closure Principle; and Ford, Bresnan and Kaplan introduced the notions of Lexical Preference and Final Arguments.

The two ideas that dominated their hypotheses and discussions were Right Association, which says roughly that postmodifiers prefer to be attached to the nearest previous possible head, and a stronger principle stipulating that argument interpretations are favored over adjunct interpretations. This latter principle is implied by Frazier and Fodor's Minimal Attachment and also by Ford, Bresnan and Kaplan's Lexical Preference.

In computational linguistics, Church, Shieber and Pereira (Church, 1980; Shieber, 1983; Pereira, 1985) proposed a shift-reduce parser for parsing English, and showed that Right Association was equivalent to preferring shifts over reductions, and that Minimal Attachment was equivalent to favoring the longest possible reduction at each point.

More recently, there have been debates, for example, between Schubert (1984, 1986) and Wilks et al. (1985), about the interaction of syntax with semantics and the role of semantics in disambiguating the classical ambiguities. Along similar lines Crain and Steedman (1984) and Hirst (1987) discuss the notion of Referential Success in which it is predicted that attachments to definite NP's would be less preferred.

In 1990 Whittemore, Ferrara, and Brunner revisited the issue of PP attachment and proposed a very successful algorithm. It describes an order in which to apply various strategies, and uses right association to resolve all conflicts. What they propose is roughly, in order: attach temporal PP's to temporal nouns or verbs, and locatives to locatives; then use lexical preference (including verbs and nouns); then use the nearest possible attachment avoiding attachment of PP's to definite NP's unless the preposition is "of"; then use right association. Their algorithm successfully accounts for all but two of the 724 tension sites in their data of type-written air travel domain dialogues.

We take it for granted that, psychologically, syntax, semantics, and pragmatics interact very tightly to achieve disambiguation. In fact, in other work (Hobbs et al., 1993), we have proposed an integrated framework for natural language processing that provides for this tight interaction. However, in this paper, we are considering

only syntactic factors. In the semantically and pragmatically unsophisticated systems of today, these are the most easily accessible factors, and even in more sophisticated systems, there will be examples that semantic and pragmatic factors alone will fail to disambiguate.

The two principles we propose may be viewed as generalizations of Minimal Attachment and Right Association.

### 3 Most Restrictive Context

The first principle might be called the Most Restrictive Context principle. It can be stated as follows:

Where a constituent can be placed in two different structures, favor the structure that places greater constraints on allowable constituents.

For example, in

John looked for Mary.

“for Mary” can be interpreted as an adverbial signaling the beneficiary of the action or as a complement of the verb “look”. Since virtually any verb phrase can take an adverbial whereas only a very few verbs can take a “for” prepositional phrase as its complement, the latter interpretation has the most restrictive context and therefore is favored.

A large number of preferences among ambiguities can be subsumed under this principle. They are enumerated below.

1. As in the above example, favor argument over adverbial interpretations for postmodifying prepositional phrases where possible. Thus, whereas in

John cooked for Mary.

“for Mary” is necessarily an adverbial, in “John looked for Mary” it is taken as a complement. Subsumable under this heuristic is the preference of “by” phrases after passives to indicate the agent rather than a location. This heuristic, together with the next type, constitutes the traditional Minimal Attachment principle. This heuristic is very strong; of 47 occurrences examined, all were in accord with the heuristic.

2. Favor arguments over mere modifiers. Thus, in

John bought a book from Mary.

the favored interpretation is “bought from Mary” rather than “book from Mary”. Where the head noun is also subcategorized for the preposition, as in,

John sold a ticket to the theater.

this principle fails to decide among the readings, and the second principle, described in the next section, becomes decisive.

This principle was surprisingly strong, but perhaps for illegitimate reasons. Of 75 potential ambiguities, all but one were in accord with the heuristic. The one exception was

HDTV provides television images with finer detail than current systems.

and even this is a close call. However, it is often very uncertain whether we should say verbs, nouns, and adjectives subcategorize for a certain preposition. For example, does “discussion” subcategorize for “with” and “about”? We are likely to say so when it yields the right parse and not to notice the possibility when it would yield the wrong parse. So our results here may not be completely unbiased.

3. Favor complement interpretations of infinitives over purpose adverbial interpretations. In

John wants his driver to go to Los Angeles.

the preferred interpretation has only the driver and not John going to Los Angeles.

Of 44 examples of potential ambiguities of this sort that we found, 41 were complements and only 3 were purpose adverbials. Even these three could have been eliminated with the simplest selectional restrictions. One example was the following

He pushed aside other business to devote all his time to this issue.

which could have been parsed analogously to

He pushed strongly all the young researchers to publish papers on their work.

A particularly intriguing example, remembering that “provide” can be ditransitive, is the following:

That is weaker than what the Bush administration needs to provide the necessary tax revenues.

4. Favor the attachment of temporal prepositional phrases to verbs or event nouns. In the preferred reading of

John saw the President during the campaign.

the seeing was during the campaign, since “President” is not an event noun. In the preferred reading of

The historian described the demonstrations during Gorbachev’s visit.

the demonstrations are during the visit. This case can be considered an example of Minimal Attachment if we assume that all verbs and event nouns have potential temporal arguments. Of 74 examples examined, 66 were in accord with this heuristic. Two that were not involved the phrase “business since August 1”.

5. Favor adverbial over object interpretations of temporal and measure noun phrases. Thus, in

John won one day in Hawaii.

“one day in Hawaii” is preferentially the time and place John won and not his prize. In

John walked 10 miles.

“10 miles” is a measure of how far he walked, not what he walked. This is an example of Most Restrictive Context because noun phrases, based on syntactic criteria alone, can always be the object of a transitive verb, whereas only temporal and measure noun phrases can function as adverbials. This case is interesting because it runs counter to Minimal Attachment. Here arguments are *disfavored*.

Of fifteen examples we found of such ambiguities, eleven agreed with the heuristic. The reason for the large percentage of examples that did not is that sports articles were among those examined, and they contained sentences like

Smith gained 1240 yards last season.

This illustrates the hidden dangers in genre selection.

6. Favor temporal nouns as adverbials over compound nominal heads. The latter interpretation is possible, as seen in

Is this a CSLI Thursday?

But the preferred reading is the temporal one that is most natural in

I saw the man Thursday.

7. Favor “that” as a complementizer rather than as a determiner. Thus, in

I know that sugar is expensive.

we are probably not referring to “that sugar”. This is a case of Most Restrictive Context because the determiner “that” can appear in any noun phrase, whereas the complementizer “that” can occur only after a small number of verbs. This is a heuristic we suspect everyone who has built a moderately large grammar has implemented, because of the frequency of the ambiguity.

8. An initial “there” is interpreted as an existential, where possible, rather than as a locative. We interpret

There is a man in the room.

as an existential declarative sentence, rather than as an utterance with an initial locative. Locatives can occur virtually anywhere, whereas the existential “there” can occur in only a very small range of contexts. Of 30 occurrences examined, 29 were in accord with the heuristic. The one exception was

There, in the midst of all those casinos, is Trump’s Taj Mahal.

The locative reading may be more common in spoken than in written discourse, however.

9. Favor predeterminers over separate noun phrases. In

Send all the money.

the reading that treats “all the” as a complex determiner is favored over the one that treats “all” as a separate complete noun phrase in indirect object position. There are very many fewer loci for predeterminers than for noun phrases, and hence this is also an example of Most Restrictive Context.

10. Favor prepositional lexical adverbs over separate adverbials. Thus, in

John did the job precisely on time.

we favor “precisely” modifying “on time” rather than “did the job”. Very many fewer adverbs can function as prepositional modifiers than can function as verbal or sentential adverbs. Of 28 occurrences examined, all but one were in accord with the heuristic. The one was

Who is going to type this all for you?

11. Group numbers with prenominal unit nouns but not with other prenominal nouns. For example, “10 mile runs” are taken to be an indeterminate number of runs of 10 miles each rather than as exactly 10 runs of a mile each. Other nouns can function the same way as unit nouns, as in “2 car garages”, but it is vastly more common to have the number attached to the head noun instead, as in “5 wine glasses”. Virtually any noun can appear as a prenominal noun, whereas only unit nouns can appear in the adjectival “10-mile” construction. Hence, for unit nouns this is the most restrictive context. While other nouns can sometimes occur in this context, it is only through a reinterpretation as a unit noun, as in “2 car garages”.

12. Disfavor headless structures. Headless structures impose no constraints, and are therefore never the most restrictive context, and thus are the least favored in cases of ambiguity. An example of this case is the sentence

John knows the best man wins.

which we interpret as a concise form of

John knows *that* the best man wins.

rather than as a concise form of

John knows the best *thing which* Man wins () .

## 4 Attach Low and Parallel

The second principle might be called the Attach Low and Parallel principle. It may be stated as follows:

Attach constituents as low as possible, and in parallel with other constituents if possible.

The cases subsumed by this principle are quite heterogeneous.

1. Where not overridden by the Most Restrictive Context principle, favor attaching postmodifiers to the closest possible site, skipping over proper nouns. Thus, where neither the verb nor the noun is subcategorized for the preposition, as in

John phoned a man in Chicago.

or where *both* the verb and the noun are subcategorized for the preposition, as in

John was given a book by a famous professor.

the noun is favored as the attachment point, since that is the lowest possible attachment point in the parse tree. This case is just the traditional Right Association.

The subcase of prepositional phrases with “of” is significant enough to be mentioned separately. We might say that every noun is subcategorized for “of” and that therefore “of” prepositional phrases are nearly always attached to the immediately preceding word. Of 250 occurrences examined, 248 satisfied this heuristic, and of the other two

Since the first reports broke of the CIA's activities, . . .

He ordered the destruction two years ago of some records.

the second would not admit an incorrect attachment in any case.

We examined 148 instances of this case not involving “of”, temporal prepositional phrases, or prepositions that are subcategorized for by possible attachment points. Of these, 116 were in accord with the heuristic and 32 were not. An example where this heuristic failed was

They abandoned hunting for food production.

For a significant number of examples (34), it did not matter where the attachment was made. For instance, in

John made coffee for Mary.

both the coffee and the making are for Mary. We counted these cases as being in accord with the heuristic, since the heuristic would yield a correct interpretation. We may refer to this phenomenon as *benign ambiguity*.

This is perhaps the place to present results on two very simple algorithms. The first is to attach prepositional phrases to the closest possible attachment point, regardless of other considerations. Of 251 occurrences examined, 125 attached to the nearest possibility, 109 to the second nearest, 14 to the third, and 3 to the fourth, fifth, or sixth. This algorithm is not especially recommended, but the results do suggest that it is safe to limit the candidate set of attachment points to the nearest two or three.

The second algorithm is to attach to the nearest possible attachment point that subcategorizes for the preposition, if there is such, assuming verbs and event nouns to subcategorize for temporal prepositional phrases, and otherwise to attach to the nearest possible attachment point. This is essentially a summary of our heuristics for prepositional phrases. Of 297 occurrences examined, this yielded the right answer on 256 and the wrong one on 41.

2. Favor prepositional readings of measure phrases over readings as separate adverbials. Thus, in

John walked 10 miles into the forest.

we preferentially take “10 miles” as modifying “into the forest” rather than “walked”, so that John is now 10 miles from the edge of the forest, rather than merely somewhere in the forest but 10 miles from his starting point. Since the preposition occurs lower in the parse tree than the verb, this is an example of Attach Low and Parallel. Note that this is a kind of “Left Association”.

3. Coordinate “both” with “and”, if possible, rather than treating it as a separate determiner. In

**John likes both intelligent and attractive women.**

the interpretation in which there are exactly two women who are intelligent and attractive is disfavored. Associating “both” with the coordinated adjectives rather than attaching it to the head noun is attaching it lower in the parse tree.

4. Distribute prenominal nouns over conjoined head nouns. In “oil sample and filter”, we mean “oil sample and oil filter”. A principle of Attach Low would not seem to be decisive in this case. Would it mean that we attach “oil” low by attaching it to “sample” or that we attach “and filter” low by attaching it to “sample”. It is because of examples like this (and the next case) that we propose the principle Attach Low *and Parallel*. We favor the reading that captures the parallelism of the two head nouns.

5. Distribute determiners and noun complements over conjoined head nouns. In “the salt and pepper on the table”, we treat “salt” and “pepper” as conjoined, rather than “the salt” and “pepper on the table”. As in the previous case, where we have a choice of what to attach low, we favor attaching parallel elements low.

6. Favor attaching adjectives to head nouns rather than prenominal nouns. We take “red boat house” to refer to a boat house that is red, rather than to a house for red boats. Like all of our principles, this preference can be overridden by semantics or convention, as in “high stress job”. Here again we could interpret Attach Low as telling us to attach “red” to “boat” or to attach “boat” to “house”. Attach Low and Parallel tells us to favor the latter.

## 5 Interaction and Overriding

There will of course be many examples where both of our principles apply. In the cases that occur with some frequency, in particular, the prepositional phrase attachment ambiguities, it seems that the Most Restrictive Context principle dominates Attach Low and Parallel. It is unclear what the interactions between these two principles should be, more generally.

These principles can be overridden by more than just semantics and pragmatics. Commas in written discourse and pauses in spoken discourse (see Bear and Price, 1990, on the latter) often function to override Attach Low and Parallel, as in

**John phoned the man, in Chicago.**

**Specify the length, in bits, of a word.**

It is the phoning that is in Chicago, and the specification is in bits while the length is of a word. Similarly, commas and pauses can override the Most Restrictive Context principle, as in

**John wants his driver, to go to Los Angeles.**

Here we prefer the purpose adverbial reading in which John and the driver both are going to Los Angeles.

## 6 Cognitive Significance

The analysis of parse preferences in terms of these two very general principles is quite appealing, and more than simply because they subsume a great many cases. They seem to relate somehow to deep principles of cognitive economy. The Most Restrictive Context principle is a matter of taking all of the available information into account in constructing interpretations. The “Low” of Attach Low and Parallel is an instance of a general cognitive heuristic to interpret features of the environment as locally as possible. The “Parallel” exemplifies a general cognitive heuristic to see similarity wherever possible, a heuristic that promotes useful generalizations.

## References

- [1] Bear, John, and Jerry Hobbs, 1988. “Localizing Expression of Ambiguity”, *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, pp. 235–241.
- [2] Bear, John, and Patti Price, 1990. “Prosody, Syntax and Parsing”, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, Pennsylvania, pp. 17–22.
- [3] Church, Kenneth, 1980. “On Memory Limitations in Natural Language Processing”, MIT Technical Report MIT/LCS/TR-245.
- [4] Crain, Stephen and Mark Steedman, 1984. “On Not Being Led Up the Garden Path: The Use of Context by the Psychological Syntax Processor,” in Dowty, Karttunen, and Zwicky (Eds.) *Natural Language Parsing*, Cambridge University Press, Cambridge.
- [5] Ford, Marylyn, Joan Bresnan, and Ronald Kaplan, 1982. “A Competence-Based Theory of Syntactic Closure,” in J. Bresnan (Ed.) *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, Massachusetts.
- [6] Frazier, Lyn and Janet Fodor, 1979. “The Sausage Machine: A New Two-Stage Parsing Model”, *Cognition* 6, pp. 291–325.
- [7] Grosz, Barbara, Norman Haas, Gary Hendrix, Jerry Hobbs, Paul Martin, Robert Moore, Jane Robinson, Stanley Rosenschein, 1982. “DIALOGIC: A Core Natural-Language Processing System”, Technical Note 270, Artificial Intelligence Center, SRI International.
- [8] Hirst, Graeme (1987) *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, New York.
- [9] Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993. “Interpretation as Abduction”, *Artificial Intelligence*, No. 63, pp. 69–142.
- [10] Kimball, John, 1973. “Seven Principles of Surface Structure Parsing in Natural Language”, *Cognition* 2(1), pp. 15–47.

- [11] Marcus, Mitchel, 1980. *A Theory of Syntactic Recognition for Natural Language*, MIT Press, Cambridge, Massachusetts.
- [12] Pereira, Fernando, 1985. “A New Characterization of Attachment Preferences,” in D. Dowty et al. (Eds.) *Natural Language Processing*, Cambridge University Press, Cambridge.
- [13] Schubert, Lenhart, 1984. “On Parsing Preferences”, *Proceedings, COLING 1984*, Stanford, California, pp. 247–250.
- [14] Schubert, Lenhart, 1986. “Are There Preference Trade-offs in Attachment Decisions?” *Proceedings, AAAI 1986*, Philadelphia, Pennsylvania, pp. 601–605.
- [15] Shieber, Stuart, 1983. “Sentence Disambiguation by a Shift-Reduce Parsing Technique”, *Proceedings, IJCAI 1983*, Washington, D.C., pp. 699–703.
- [16] Wanner, Eric, and Michael Maratsos, 1978. “An ATN Approach to Comprehension,” in Halle, Bresnan, and Miller (Eds.) *Linguistic Theory and Psychological Reality*. MIT Press, Cambridge, Massachusetts.
- [17] Whittemore, Greg, Kathleen Ferrara, and Hans Brunner, 1990. “Empirical Study of Predictive Powers of Simple Attachment Schemes for Post-modifier Prepositional Phrases”, *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, Pennsylvania, pp. 23–30.
- [18] Wilks, Yorick, Xiuming Huang, and Dan Fass, 1985. “Syntax, Preference and Right Attachment”, *Proceedings, IJCAI 1985*, Los Angeles, California, pp. 779–784.

# Varieties of Heuristics in Sentence Parsing\*

Makoto Nagao

Dept. of Electrical Engineering, Kyoto University

Yoshida-honmachi, Sakyo, Kyoto, 606, Japan

e-mail: [nagao@kuee.kyoto-u.ac.jp](mailto:nagao@kuee.kyoto-u.ac.jp)

## 1 Kinds of Grammars and Their Characteristics

There are many methods of sentence parsing, but parsing always presupposes a grammar, which is usually composed of a set of so-called grammatical rules or rewriting rules. There are many grammars proposed so far, and many parsing algorithms have been developed based on these grammars. Characteristics of these parsing algorithms are a direct reflection of the features of the grammar formalisms used by sentence parsing, so that we have to clarify the basic characteristics of these grammars.

We can classify grammars so far proposed into the following few classes:

(i) Phrase Structure Grammar (PSG)

Context-free PSG, Context-sensitive PSG, Augmented Transition Network Grammar, Definite Clause Grammar, Categorial Grammar, Lexical Functional Grammar, Generalized PSG, Head Driven PSG, Tree Adjoining Grammar, ...

(ii) Dependency Grammar

(iii) Case Grammar

(iv) Systemic Grammar

(v) Montague Grammar

In the following we will discuss the basic ideas behind these grammars, and compare them contrastively from the standpoint of parsing.

### 1.1 Phrase Structure Grammar

Phrase structure grammar was proposed by N. Chomsky from the standpoint of sentence generation. This means that this grammar formalism is not necessarily fitted to the analysis of sentences. This will become clear when we consider the meaning of a

---

\*This paper was presented at the 3rd International Workshop on Parsing Technologies, Tilburg/Durbuy, 1993, 8, 10–13, as an invited talk, but was not included in the proceedings.

grammar rule such as  $S \rightarrow NP \cdot VP$ . This rule declares that a sentence *should* be composed of NP followed by VP, or that a sentence *presupposes* the existence of NP followed by the existence of VP. This means that a sentence which is outside of this definition is excluded from the scope of the language. In this way this grammar formalism gives the definition of a language.

There is always a gap between an existing language and the set of sentences which a grammar can produce. The gap is not small, but actually very big. Sentences which we speak or write are essentially free, and cannot be grasped by such an artificial framework. We always encounter sentences or expressions which cannot be explained by a grammar, and we are forced to improve or add rewriting rules constantly. When a grammar is proposed as a tool to give a conceptual explanation of sentential structures of a language to a human being, a simple basic grammar will be sufficient. But when a grammar is to be used by a computer to parse existing sentences mechanically, there must be a very precise grammar, and its constant improvement will be required.

Japanese language is quite different from English and many other European languages. Japanese language is more free in word order change than these other languages, and there are varieties of omissions of essential components in a Japanese sentence. PSG has difficulty in handling these phenomena because the grammar formalism presupposes the word (phrase) order as is specified in grammar rules. Basically this grammar formalism does not fit languages which have free word order and where the notion of phrase structure does not hold.

Similar discussion holds for ellipsis. When PSG is used for the analysis of a language which has varieties of ellipses, we have to try rule applications not only of the rules which have every component, but also of the rules which do not have (ignore) some of these components because it is quite difficult to specify under what condition a certain element can be omitted. This is almost impossible to execute. And the concept of grammatical restriction will not hold in such a case; that is, the phrase structure grammar will have no meaning any more.

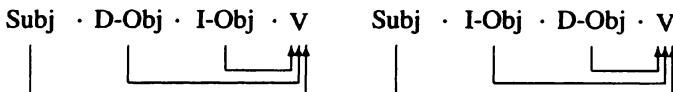
PSG has such a serious problem in the analysis of sentences of at least a certain kind of languages such as Japanese. Nevertheless, many people use PSG for parsing without such considerations. We must see that there are other grammar formalisms which may be more suitable for sentence parsing.

## 1.2 Dependency Grammar

Dependency grammar (DG), which was treated in detail first by L. Tesnière is a grammar which is known as “Kakariuke” grammar in Japanese and which has been very popular for more than fifty years in Japan. This grammar is totally different from PSG in the sense that while PSG is a kind of language definition tool, DG is a kind of interpretation tool of a given sentence. DG clarifies which word modifies or depends on which other word. It constructs the modifier-modifiee relations between the words in a sentence, and this is always possible because a word in a sentence always has a relation to another word in the same sentence. It does not presuppose anything, but just clarifies the word relations in a sentence. In this sense DG can be regarded as a grammar for interpretation.

The interpretation power of this grammar, however, is far weaker than that of PSG. DG does not say anything about subject, object, etc., but just say that this word modifies

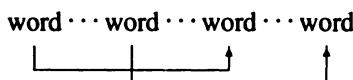
that, etc. However, this weak property becomes profitable, for example, for word-order change in Japanese because the modifier-modifiee relation is not influenced by the word-order change, as shown in the following:



DG is not concerned with the omitted words, and the analysis process is always the same.

So far so good. But from the standpoint of machine translation or some other natural language processing, the role of each word or phrase in a sentence must be clarified more accurately. In Japanese this is shown approximately by the information from verb conjugation, postpositions attached to nouns, and sometimes by the position of the word. So some additional analysis should follow DG analysis.

A serious problem in DG is an ambiguity problem—that a word can modify two or more words which follow the word in a sentence—and the decision is difficult. To solve this problem we have to prepare a good dictionary where the degree of affinity of two words is described. This is, however, not a particular problem for DG alone. It is a common problem for almost all grammars including PSG. This problem is relieved to a certain extent by the so-called non-crossing condition in the Japanese language, which can be illustrated by the following disallowed situation:



This condition is useful for the elimination of redundant checks of modifier-modifiee relation in Japanese.

### 1.3 Case Grammar

Case grammar (CG), which is proposed by C. Fillmore, is quite different from the above two grammar formalisms in the point that CG aims to clarify the roles of words in a sentence from the standpoint of meaning or of conceptual function of a word to a predicate. Therefore CG does not care about the word order in a sentence.

Case grammar interpretation of a sentence is represented by a case frame whose slots are filled in by words in a sentence. This case frame representation can be seen as a meaning representation of a sentence, so that it is comparatively neutral to a language. That is, sentences in different languages which express the same contents may have the same case frame. This is the main reason why many machine translation systems have adopted the case frame representation as the final goal of the sentence analysis and the starting point of sentence generation.

One of the difficulties of CG is that we have to tackle the difficult problem of formalizing the meaning representation because the grammar is based on meaning. Fillmore did not explicitly discuss this problem because was not interested in machine processing but he relied on the human ability to interpret meaning. What he and his followers discussed seriously was what kinds of cases (slot functions in case frame) must be set up, such as agent, object, instrument, etc. What we, natural language processing researchers, had to do was to establish a semantic marker system which is powerful enough to distinguish different usages of verbs and nouns. Case frame must specify what kinds of nouns can be an agent, object, instrument, etc. of a particular verb in particular usage or meaning. The verb "kakeru" in Japanese, for example, has more than thirty different usage patterns, and the semantic markers should be detailed enough to be able to select the correct nouns for these usage patterns. It has been made clear by the effort of NTT researchers that 2-3 thousand semantic markers are necessary for the satisfactory description of every different usage of verb patterns (Ikehara 1991).

## 1.4 Other Grammar Formalisms

There are many grammar formalisms which can be classified into PSG, and these have the same problems which were discussed in Section 1.1. Montague grammar is based on formal logic, and the meaning of a sentence is represented by a logical formula. Present-day formal logic has a limited power of representation and it is impossible to express rich information of natural language in its limited formalism. There were several attempt at machine translation based on a Montague representation but these all failed because of the poor expressive power of logic and also of the difficulty of transforming a sentential expression to a proper logical form.

Systemic grammar, which M.A.K. Halliday proposed, is unique in the sense that it distinguishes three components in a sentence, namely ideational function, textual function and interpersonal function. Japanese language is rich in interpersonal functions. It has a sophisticated honorific expression system, for example, which uses specific words that affect sentential styles. Textual function corresponds roughly to discourse or contextual function. Systemic grammar is developed essentially for the generation of a sentence, and it is not clear whether it is useful to the analysis of a sentence.

# 2 Varieties of Heuristic Components in Parsing Sentences

## 2.1 Ambiguity Resolution in Sentence Parsing

In sentence parsing we are always confronted with the problem of ambiguity resolution of modifier-modifiee relation. To solve this problem we need such information as

- (i) semantic consistency of modifier and modifiee.
- (ii) word order and distance between modifier and modifiee. What kinds of words or symbols exist in between these two words are important for the rejection of the modifier-modifiee relation between the two words.
- (iii) consistency with contextual/situational information.

(i) is checked by the consistency of semantic markers. To realize this all the words in a dictionary must be given proper semantic markers for their distinctive meanings. We may be able to utilize inclusion relations of semantic markers for the consistency checking.

(ii) is not easy to handle. Basically the modifier modifies the nearest possible modifiee, but it is not always true. We cannot specify a semantic relation so exactly as to accept just correct ones and reject the others. There always exist word pairs which cannot be accepted or rejected definitely. In these cases we have to look for other information to determine acceptance or rejection. It is often very useful to see words or symbols in between the two candidate words in a Japanese sentence. For example, the first word of a sentence which has -wa as a suffix usually modifies the last predicate of the sentence. But there are several cases where this condition does not hold. The following is an example:

Tokyo-wa    bukka-ga    takai-ga,    inaka-wa    yasui.  
(Tokyo)    (price)    (expensive)    (countryside)    (cheap)  
(Things are expensive in Tokyo, but cheap in the countryside.)

In this sentence there is another -wa and a predicate takai in between Tokyo-wa and the final predicate yasui, and these prevent the relation of these two words. In the following sentences,

Tokyo-wa    bukka-ga    takai-ga    hito-ga    atsumaru.  
Tokyo-wa,    bukka-ga    takai-ga    hito-ga    atsumaru.  
(human being)    (gather)

the first sentence permits two interpretations, that is, Tokyo-wa modifies either takai or atsumaru (gather). But the second usually has one interpretation, that is, Tokyo-wa modifies atsumaru. This is caused by the comma after Tokyo-wa. We have developed a sophisticated algorithm to solve these modifier-modifiee problems (Kurohashi and Nagao 1992a).

As for case (iii) we have a famous example,

I saw a woman in the garden with a telescope.

Correct interpretation is possible only when we know the real situation of the utterance.

## 2.2 Anaphora and Ellipsis

When we proceed the analysis of a sentence from morphological analysis, syntactic analysis and semantic interpretation, that is, the transformation to case frame representation, anaphora and ellipsis are handled usually at the last stage of case frame representation. We can recognize that there is an ellipsis when we find out a vacant slot in a case frame representation of a sentence. We may be able to infer and recover this by utilizing contextual information so far obtained (sometimes we have to see some more words or sentences after this ellipsis position). Inference and recovery of a proper word is not easy, but we can write varieties of inference rules which check grammatical and semantic information requested from the case slot default information and which utilize the contextual information so far obtained.

As for the anaphora resolution, we can utilize grammatical information from pronominal words and write similar heuristic rules as above. We can of course think of another analysis process where anaphora determination is done just before the case frame transformation.

Looking for a proper word or concept for a pronominal reference or an ellipsis is a problem of language understanding. We have an example like the following.

Merikenko to Satou-o yoku maze, atatameru.  
(flour) (and) (sugar) (well) (mix) (warm up)

where there is an omission of an object of the predicative verb, atatameru. When we trace back the sentence we encounter the nouns, Satou and Merikenko. These are the candidates for the omitted object. However, the actual thing to warm up is neither of the two. It is the mixture of these two materials, which does not appear explicitly in the sentence. We must introduce an inference mechanism that a mixture is created when two materials are “Mazeru-ed” (mixed). We have to have a mechanism to understand the meaning of a sentence and do the action which the sentence specifies, and get the result by applying an inference rule. In this way when there is a series of sentences which describe actions performed in time sequence, each action is supposed to be applied to the object which is produced as the result of the previous actions. We have to write many heuristic rules to produce these objects.

We may be forced to infer more than this. For the above example we cannot warm up the mixture of flour and sugar directly. The mixture must be in a certain thing such as a bowl. The determination of pronominal reference and the recovery of omission are quite difficult. We have to clarify what the language understanding is and what the common sense reasoning is.

## 2.3 Referential Property and Number

Besides the determination of pronominal reference and ellipsis we have to clarify the referential property (definite, indefinite, generic) and number (singular, plural, uncountable) of a noun in a sentence. In the Japanese language there are no indications such as articles and number suffixes in English to show these properties of a noun. Therefore we must estimate the referential property and number of a noun in a Japanese sentence. This requires language understanding by contextual information. It is difficult to achieve this at the present stage of research.

On the other hand we have to check how much information we can get to infer these properties of a noun from the sentence in which the noun appears. We are interested in this problem and tried to construct a kind of expert system to solve this problem (Murata and Nagao 1993). We wrote 84 heuristic rules to infer the referential property of a noun, and 48 heuristic rules to infer the number property.

Let us consider the following example.

Kinou katta piano-no ichidai-wa choritsu-ga yokunai.  
(yesterday) (bought) (piano) (one unit) (tuning) (no good)  
(One of the pianos which I bought yesterday is not tuned well.)

Piano in this sentence is modified by kinou katta (piano which I bought yesterday). So this piano is a concrete object which I have now, and the noun “piano” has the property definite. Piano is followed by ichidai, which means “one of ···”, so that we can infer that “piano” is not singular. Many such heuristic rules are written in the form of expert system rules.

The test of this system was done for two different sample sentence sets. The first test was done for the nouns in a set of sentences which were referenced in the process of writing heuristic rules. The success rate were 85.5% for the referential property and 89.0% for the number property. The second test was done for the nouns in a set of newly given sentences. The success rates were 68.9% and 85.6% for the referential property and the number property, respectively. This kind of analysis of nouns in a sentence is very important when we consider the construction of better machine translation systems.

## 2.4 Tense, Aspect, and Modality

We have to write many heuristic rules to interpret correctly the tense, aspect and modality of a predicate. These are particularly important in the interpretation of tense, aspect and modality of two or more predicates in a sentence when one of these is in an (i) embedded sentence, (ii) subordinate clause, or (iii) coordination. Many standard grammar books explain these relations in detail, but it is still very difficult to write expert system rules precisely for these relations.

Let us consider a pair of languages, for example English and Japanese. The categories and properties of tense, aspect and modality of Japanese are completely different from those of English. For example the Japanese language has no present perfect tense. This tense is often expressed by adverbs in Japanese such as follows.

Kare-wa ima tuita.  
(he) (now) (arrived)  
(He has just arrived.)

A more difficult situation exists where there are no such adverbs and we have to infer the tense from the situation.

Kare-wa Kyoto-e kita. Soshite ima kankou-o shiteiru.  
(he) (Kyoto) (came) (and) (now) (sight seeing) (do)  
(He has come to Kyoto. And he is now doing sight seeing.)

Kare-wa Kyoto-e kita. Soshite Kyoto daigaku-o sotsugyo shita.  
(he) (Kyoto) (came) (and) (Kyoto University) (graduate) (did)  
(He came to Kyoto. And he graduated Kyoto University.)

Sometimes different tense expressions in Japanese correspond to the same tense in English. The following is an example,

Sore-wa 10 nen mae-no koto-de aru.  
(It) (10 years) (before) (matter) (is)  
Sore-wa 10 nen mae-no koto-de atta.  
(was)  
(It was 10 years before.)

All of these problems are related to the problem of discourse. We don't know how many such problems there are and how many expert system rules we have to write. The first problem we have to tackle will be to clarify how many different categories of discourse problems there are.

## 3 Stage Design for Parsing Sentences

### 3.1 Step-by-Step Analysis

As was mentioned in Section 1 each grammar formalism has its own characteristics, and so we have to choose carefully the best grammar formalism for a particular purpose. We believe that the following steps will be the best for the analysis of Japanese sentences for machine translation (Kurohashi and Nagao 1993).

- (i) morphological analysis
- (ii) detection of parallel structures
- (iii) dependency analysis
- (iv) case frame analysis
- (v) textual function analysis
- (vi) interpersonal function analysis

We will be able to design the analysis stage in different ways, such that the dependency analysis is skipped and the case frame is obtained from the result of morphological analysis, or that the total process may be merged into one by the constraint programming methodology and the final result is to be produced from the input sentence. However we don't recommend such processes. We believe that the best way is to divide the whole process into many subprocesses, and to perform small transformations at each subprocesses. The reason is that the analysis is an information losing process and it must be done very carefully in small steps so that important information will not be lost by a drastic change. There is an additional advantage that when the process is divided into many stages we can understand the details of each stage more easily and we can do the improvement of each stage independently.

### 3.2 Detection of Parallel Structures

Almost all the current parsing systems fail in the analysis of a long sentence, for example, a sentence composed of more than thirty English words or more than seventy Japanese characters. Nobody has analyzed the reasons for this difficulty. Probably very many factors are mixed up and the combinations are enormous, all of which are equally possible in causing errors in the analysis. There is no one major reason for the failure. We have to improve almost all the parts of the parsing process, particularly grammar rules.

On the other hand, we have to pose a question: Why is a sentence so long? The main reason is that people write a long sentence by connecting phrases, clauses and

sentences into one. Therefore an important point in the analysis of a long sentence is to find out such conjunctions. Serious efforts have been done to this problem so far by writing many grammar rules which check something like semantic similarities which may exist in the head nouns or main verbs of conjunctive structures. But there has been no significant improvement so far. This indicates that we have to think about a completely new approach to finding parallel structures in a long sentence.

What we have recently developed is based on the assumption that parallel phrases/clauses /sentences will have certain similarities in the sequence of words and their grammatical structures as a whole. We had to compare two word-strings of arbitrary lengths from these aspects and to get an overall similarity value. Because we had to compare word-string pairs of arbitrary lengths in a sentence we adopted the dynamic programming method to calculate overall similarity values for all the possible word-string pairs of arbitrary lengths and to get the best one (the details are given in the paper Kurohashi and Nagao 1992b). The result was unexpectedly good. This algorithm has achieved more than 90% success rate in finding parallel structures of different kinds, and many long Japanese sentences which were composed of more than 100 characters were successfully analyzed.

The idea behind this algorithm is just to find out similar word strings in a sentence, which has nothing to do with the existing notion of grammatical rules. It is closer to human cognitive action which is vague, but which is very reliable in the global recognition process. The human brain may not work like the application of grammatical rules, but may work like the similarity detection mentioned here.

This algorithm was inserted in between the morphological analysis and the dependency analysis in our parsing system, and achieved a very good performance in sentence parsing.

### **3.3 Dependency Analysis and Conversion to Case Frame Representation**

After the detection of conjunctive structures, the dependency analysis is first done to these parts, and then to the whole sentential structure. By adding this parallel structure detection algorithm the accuracy of the dependency analysis has become very high.

The transformation of the dependency tree of a sentence to a case frame representation is not difficult because the noun-verb modifying relations have been obtained at the stage of the dependency analysis. The work to do at this stage is to check in which case slots those nouns will come (Kurohashi and Nagao 1993). If there remains a vacant case slot after the assignment of words to suitable case slots, it is judged as an omission.

### **3.4 Recognition of a Phrase Unit**

People utter a sentence not word by word, but phrase by phrase. This phrase unit is what Margaret Masterman called a breathgroup. This phrase unit has a unique meaning although each word which is a component of a phrase may have several meanings. This phrase unit is quite different from that of PSG in the sense that a phrase in PSG is hierarchically recursive, but a phrase here is not.

Example-based machine translation, which has been recognized as giving better quality translation than the other methods, memorizes lots of example phrases and

their translations as pairs, and a target language sentence is composed of the phrasal translations. Example phrases in this case are just this phrase unit which corresponds to the breathgroup. There is a discussion in machine translation study about the translation equivalence unit. Words have many meanings and are very ambiguous, so that they are not good for the translation equivalence unit. Sentences have structural ambiguity and are also too big to be an equivalence unit. Therefore phrases are good candidates as translation equivalence unit.

Nowadays there are lots of efforts to collect typical example phrases, and to construct a phrasal dictionary, because this dictionary contributes very much for the improvement of translation quality in machine translation. So recognizing phrases properly in a sentence has become an important task, particularly in example-based translation. When the phrases in a sentence are correctly recognized the analysis of a whole sentence becomes easy because the relations among these phrases are not so difficult to determine.

## 4 Conclusion

We have discussed the essential properties of different grammar formalisms, and suggested that a proper grammar must be chosen for a particular purpose. For example, generative grammar is not suitable for sentence analysis, so that the idea of bi-directional grammar, which aims at using the same grammar for analysis and synthesis of a sentence in machine translation, is to be reconsidered.

A sentence includes lots of information, such as syntactic information, semantic information, textual or rhetorical information, interpersonal information, and so on. We do not have any formal methods to detect this information. Even at the level of syntactic information PSG and other grammar formalisms are just a framework to explain basic structures of a language. No one knows how the human brain works to speak and recognize a language. We can just approximate the human language mechanism to a certain extent from various aspects. All are heuristic. Grammar formalisms, even PSG, are a kind of expert system. We have to write many expert systems to approximate a language at the levels of morphology, syntax, semantics, discourse, etc. Therefore we have to consider a parallel or pipeline execution structure of these expert systems because the total system is too big and complex to be executed sequentially. This will be a very interesting software problem in the near future.

## References

- [1] Ikehara, S., et al., "Semantic Analysis Dictionaries for Machine Translation", *IPSJ-NLP 84-13*, July 19, 1991 (in Japanese).
- [2] Kurohashi, S. and M. Nagao, "A Syntactic Analysis Method of Long Japanese Sentences based on Conjunctive Structures' Detection", *IPSJ-NLP 88-1*, March 12, 1992 (in Japanese).
- [3] Kurohashi S. and M. Nagao, "Dynamic Programming Method for Analyzing Conjunctive Structures in Japanese", In *Proc. of COLING '92*, July 23-28, 1992.

- [4] Kurohashi S. and M. Nagao, "Structural Disambiguation in Japanese by Evaluating Case Structures based on Examples in Case Frame Dictionary", In *Proc. of IWPT '93*, August 10-13, 1993.
- [5] Murata, M. and M. Nagao, "Determination of referential property and number of nouns in Japanese sentences for machine translation into English", In *Proc. of TMI '93*, July 14-16, 1993.

# UD, yet another unification device\*

R. Johnson, IDSIA, Lugano  
M. Rosner, IDSIA, Lugano  
*e-mail: mike@idsia.uu.ch*

## Abstract

This article<sup>1</sup> describes some of the features of a sophisticated language and environment designed for experimentation with unification-oriented linguistic descriptions. The system, called UD, has to date been used successfully as a development and prototyping tool in a research project on the application of situation schemata to the representation of real text, and in extensive experimentation in machine translation.

While the UD language bears close resemblances to all the well-known unification grammar formalisms, it offers a wider range of features than any single alternative, plus powerful facilities for notational abstraction which allow users to simulate different theoretical approaches in a natural way.

After a brief discussion of the motivation for implementing yet another unification device, the main body of the article is devoted to a description of the most important novel features of UD.

## 1 Introduction

The development of UD arose out of the need to have available a full set of prototyping and development tools for a number of different research projects in computational linguistics, all involving extensive text coverage in several languages: principally a demanding machine translation exercise and a substantial investigation into some practical applications of situation semantics (Rupp, Johnson and Rosner, 1992).

The interaction between users and implementers has figured largely in the development of the system, and a major reason for the richness of its language and environment has been the pressure to accommodate the needs of a group of linguists working on three or four languages simultaneously and importing ideas from a variety of different theoretical backgrounds.

---

\*We thank Suissestra and the University of Geneva for supporting the work reported in this article, and the ACL for granting reproduction rights. We are grateful to all our former colleagues in ISSCO, and to all UD users for their help and encouragement. Special thanks are due to C.J. Rupp for being a willing and constructive guinea-pig, as well as for allowing us to plunder his work for German examples.

<sup>1</sup>This article is a slightly updated version of the authors' "A rich environment for experimentation with unification grammars" that appeared in the Proceedings of ACLE-89, Manchester. At the time of publication, the novelty of the system lay in the fact that it provided a number of experimental features, as described here, in an implementation that was not only freely available but also *efficient*, even by today's standards.

Historically UD evolved out of a near relative of PATR-II (see Shieber, 1984) and its origins are still apparent, not least in the notation. In the course of development, however, UD has been enriched with ideas from many other sources, most notably from LFG (Bresnan, 1982) and HPSG (Sag and Pollard, 1987).

Among the language features mentioned in the article are

- a wide range of data types, including lists, trees and user-restricted types, in addition to the normal feature structures;
- comprehensive treatment of disjunction;
- dynamic binding of pathname segments.

A particular article of faith which has been very influential in our work has been the conviction that well-designed programming languages (including ones used primarily by linguists), should not only supply a set of primitives which are appropriate for the application domain but should also contain *within themselves* sufficient apparatus to enable the user to create new abstractions which can be tuned to a particular view of the data.

We have therefore paid particular attention to a construct which in UD we call a *relational abstraction*, a generalisation of PATR-II templates which can take arguments and which allow multiple, recursive definition. In many respects relational abstractions resemble Prolog procedures, but with a declarative semantics implemented in terms of a typical feature-structure unifier.

## 1.1 Structure of the article

Section 2 gives a concise summary of the semantics of the basic UD unifier. This serves as a basis for an informal discussion, in Section 3, of our implementation of relational abstractions in terms of ‘lazy’ unification. The final section contains a few remarks on the issue of completeness, and a brief survey of some other language features.

## 2 Basic Unifier Semantics

In addition to the usual atoms and feature structures, the UD unifier also handles lists, trees, typed instances, and positive and negative disjunctions of atoms. This section contains the definition of unification over these constructs and employs certain notational conventions to represent these primitive UD data types, as shown in figure 1.

Throughout the description, the metavariables  $U$  and  $V$  stand for objects of arbitrary type. Three other special symbols are used:

Type name	Notation
atom	$ABC$
list	$[U V]$
n-ary tree	$V_0(V_1, \dots, V_n)$
+ve disjunction	$/C_1, \dots, C_r/$
-ve disjunction	$\neg/C_1, \dots, C_r/$
feature structure	$\{< A_1, V_1 >, \dots, < A_r, V_r >\}$
typed instance	$< C, \{< A_1, V_1 >, \dots, < A_n, V_n >\} >$

Figure 1: Notational Conventions

- [1]  $\sqcup$  is commutative:  $U \sqcup V = V \sqcup U$
- [2]  $\top$  is the identity:  $V \sqcup \top = V$
- [3]  $\sqcup$  is  $\perp$ -preserving:  $V \sqcup \perp = \perp$

Figure 2: Algebraic Properties

$\sqcup$  stands for the unification operator

$\top$  stands for top, the underdefined element.

$\perp$  stands for bottom, the overdefined element that corresponds to failure.

The semantics of unification proper are summarised in figures 2–5: Clauses [1]–[3] define its algebraic properties; clauses [4]–[6] define unification over constants, lists and trees.

In figure 4, clause [7] treats positive and negative disjunctions with respect to sets of atomic values. In figure 5, clause [8] deals with feature structures and *typed instances*. Intuitively, type assignment is a method of strictly constraining the set of attributes admissible in a feature structure.

Any case not covered by [1]–[8] yields  $\perp$ . Moreover, all the complex type constructors are strict, yielding  $\perp$  if applied to any argument that is itself  $\perp$ .

The extensions to a conventional feature structure unifier described in this section are little more than cosmetic frills, most of which could be simulated in an orthodox, PATR-style environment, even if with some loss of descriptive clarity.

In the rest of the article, we discuss a further enhancement which dramatically and perhaps controversially extends the expressive power of the language.

### 3 Extending the Unifier

The major shortcoming of typical PATR-style languages is their lack of facilities for defining new abstractions and expressing linguistic generalisations not foreseen (or even foreseeable) by the language designer. This becomes a serious issue when, as in our

- [4] constants:  $C_1 \sqcup C_2 = C_1$ , if  $C_1 = C_2$
- [5] lists:  $[U_1|U_2] \sqcup [V_1|V_2] = [U_1 \sqcup V_1|U_2 \sqcup V_2]$
- [6] trees:  $U_0(U_1, \dots, U_n) \sqcup V_0(V_1, \dots, V_n) = U_0 \sqcup V_0(U_1 \sqcup V_1, \dots, U_n \sqcup V_n)$

Figure 3: Unification of constants, lists and trees

- [7]  $/C_1, \dots, C_n/ \sqcup C = C$   
if  $C \in \{C_1, \dots, C_n\}$
- $/A_1, \dots, A_p/ \sqcup /B_1, \dots, B_q/ = /C_1, \dots, C_r/,$   
if  $C_i \in \{A_1, \dots, A_p\}$  and  $C_i \in \{B_1, \dots, B_q\}$ ,  
 $1 \leq i \leq r$ , provided (for  $r > 0$ )
- $\neg/C_1, \dots, C_n/ \sqcup C = C,$   
if  $C \neq C_i$ ,  $1 \leq i \leq n$
- $\neg/A_1, \dots, A_p/ \sqcup \neg/B_1, \dots, B_q/ = \neg/C_1, \dots, C_r/,$   
where  $C_i \in \{A_1, \dots, A_p\}$  or  $C_i \in \{B_1, \dots, B_q\}$ ,  
 $1 \leq i \leq r$
- $/A_1, \dots, A_p/ \sqcup \neg/B_1, \dots, B_q/ = \neg/C_1, \dots, C_r/,$   
where  $C_i \in \{A_1, \dots, A_p\}$  and  $C_i \notin \{B_1, \dots, B_q\}$ ,  
 $1 \leq i \leq r$

Figure 4: Unification of +ve and -ve atomic value disjunctions

[8]

$$\begin{aligned}
 & \{< A_1, U_1 >, \dots, < A_p, U_p >\} \sqcup \\
 & \{< B_1, V_1 >, \dots, < B_q, V_q >\} = \\
 & \{< A_i, U_i > | A_i \notin \{B_1, \dots, B_q\}\} \cup \\
 & \{< B_j, U_j > | B_j \notin \{A_1, \dots, A_p\}\} \cup \\
 & \quad \{< A_i, U_i \sqcup V_j > | A_i = B_j\} \\
 & \quad 1 \leq i \leq p, 1 \leq j \leq q
 \end{aligned}$$

$$\begin{aligned}
 < C, \{< A_1, U_1 >, \dots, < A_p, U_p >\} > \sqcup \\
 < C, \{< A_1, V_1 >, \dots, < A_p, V_p >\} > = \\
 < C, \{< A_1, U_1 \sqcup V_1 >, \dots, < A_p, U_p \sqcup V_p >\} >
 \end{aligned}$$

$$\begin{aligned}
 < C, \{< A_1, U_1 >, \dots, < A_p, U_p >\} > \sqcup \\
 & \{< B_1, V_1 >, \dots, < B_q, V_q >\} = \\
 < C, \{< A_i, U_i > | A_i \notin \{B_1, \dots, B_q\}\} \cup \\
 & \quad \{< A_i, U_i \sqcup V_j > | A_i = B_j\} >, \\
 & \quad 1 \leq i \leq p, 1 \leq j \leq q \\
 & (\text{for } \{B_1, \dots, B_q\} \subseteq \{A_1, \dots, A_p\})
 \end{aligned}$$

Figure 5: Unification of feature structures and typed instances

own case, quite large teams of linguists need to develop several large descriptions simultaneously.

To meet this need, UD provides a powerful abstraction mechanism which is notationally similar to a Prolog *procedure*, but having a strictly declarative interpretation. We use the term *relational abstraction*<sup>2</sup> to emphasise the non-procedural nature of the construct.

### 3.1 Some Examples of Relational Abstraction

The examples in this section are all adapted from a description of a large subset of German written in UD (Rupp, 1990). As well as relational abstractions, two other UD features are introduced here: a built-in list concatenation operator ‘++’ and generalised disjunction, notated by curly brackets (e.g. {X,Y}). These are discussed briefly in Section 4.

The first example illustrates a relation `Merge`, used to collect together the semantics of an arbitrary number of modifiers in some list `X` into the semantics of their head `Y`. Its definition in the external syntax of the current UD version is

```
Merge(X, Y) :  
    !Merge-all(X, <Y desc cond>, <Y desc ind>)
```

(The invocation operator ‘!’ is an artefact of the LALR(1) compiler used to compile the external notation - one day it will go away. `X` and `Y` should, in this context, be variables over feature structures. The `desc`, `cond` and `ind` attributes are intended to be mnemonics for, respectively, ‘description’, (a list of) ‘conditions’ and ‘indeterminate’.)

`Merge` is defined in terms of a second relation, `Merge-all`, whose definition is

```
Merge-all([Hd|T1], <Hd desc cond> ++ L, Ind) :  
    Ind = <Hd desc ind>  
    !Merge-all(T1, L, Ind)
```

```
Merge-all([], [], Ind)
```

`Merge-all` does all the hard work, making sure that all the indeterminates are consistent and recursively combining together the condition lists.

Although these definitions look suspiciously like pieces of Prolog, to which we are clearly indebted for the notation, the important difference, which we already referred to above, is that the interpretation of `Merge` and `Merge-all` is strictly declarative.

The best examples of the practical advantages of this kind of abstraction tend to be in the lexicon, typically used to decouple the great complexity of lexically oriented descriptions from the intuitive definitions often expected from dictionary coders. As illustration, without entering into discussion of the underlying complexity, for which we

---

<sup>2</sup>Relational abstractions are comparable, for example, in spirit and in syntax, to parametric sorts in the CUF (see D’orre and Eisele, 1991).

unfortunately do not have space here, we give an external form of a lexical entry for some of the senses of the German verb “tr'aumen”.

This is a real entry taken from an HPSG-inspired analysis mapping into a quite sophisticated situation semantics representation. All of the necessary information is encoded into the four lines of the entry; the expansions of Pref, Loctype and Subcat are all themselves written in UD. The feature -prefix is a flag interpreted by a separate morphological component to mean that “tr'aumen” has no unstressed prefix and can take ‘ge-’ in its past participle form.

```
traeumen -prefix
!Pref(none)
!Loctype([project])
!Subcat(np(nom), {vp(inf,squi), pp(von,dat)})
```

Pref is a syntactic abstraction used in unraveling the syntax of German separable prefixes. Loctype is a rudimentary encoding of *Actionsart*.

Subcat contains all the information necessary for mapping instances of verbs with VP or PP complements to a situation schema (Fenstad, Halvorsen, Langholm and van Benthem, 1987; Rupp, Johnson and Rosner, 1992). Here, for completeness but without further discussion, are the relevant fragments of the definition of Subcat.

```
Subcat(np(nom), pp(P,C)) :
!Normal
!Obl(Pobj,P,C,X)
!Arg(X,2)
<* subcat> = [Pobj|T]
!Assign(T,__)

Subcat(np(nom), vp(F,squi)) :
!ControlVerb
!Vcomp(VP,F,NP,Sit)
!Arg(Sit,2)
<* subcat> = [VP|T]
!Assign(T,X)
F = inf/bse
!Control(X,NP)

Assign([X], X) :
<* voice> = active
!Subj(X)
!Arg(X,1)

Assign({[Y]}, [], Z) :
<* voice> = passive
<* vform> = psp
!Takes(none)
!Obl(Y,von,dat,Z)
!Arg(Z,1)
```

## 4 Implementation of the Extensions

In this section we describe briefly the algorithm used to implement a declarative semantics for relational abstractions, concluding with some remarks on further interesting extensions which can be implemented naturally once the basic algorithm is in place. For the moment, we have only an informal characterisation, but a more formal treatment is in preparation (Johnson and Rupp, forthcoming).

### 4.1 The solution algorithm

The main problem which arises when we introduce relational abstractions into the language is that some unifications which would ultimately converge may not converge locally (i.e. at some given intermediate stage in a derivation) if insufficient information is available at the time when the unification is attempted (of course some pathological cases may not converge at all—we return to this question below).

We cope with this by defining an argument to the unifier as a *pair*  $\langle I, K \rangle$ , consisting of an *information structure*  $I$  belonging to one of the types listed in section 2, plus an *agenda*  $K$  which holds the set of as yet unresolved constraints potentially holding over  $I$ . Unification of two objects,

$$\langle I_1, K_1 \rangle \sqcup \langle I_2, K_2 \rangle$$

involves the attempt to resolve the pooled set of constraints  $K_1 \cup K_2 = K_0$  with respect to the newly unified information structure  $I_0 = I_1 \sqcup I_2$ , if it exists.

The question of deciding whether or not some given constraint set will converge locally is solved by a very simple heuristic. First we observe that application of the constraint pool  $K_0$  to  $I_0$  is likely to be non-deterministic, leading to a *set* of possible solutions. Growth of this solution set can be contained locally in a simple way, by constraining each potentially troublesome (i.e. recursively defined) member of  $K_0$  to apply only once for each of its possible expansions, and freezing possible continuations in a new constraint set.

After one iteration of this process we are then left with a set of pairs  $\{\langle J_1, L_1 \rangle, \dots, \langle J_r, L_r \rangle\}$ , where the  $L_i$  are the current constraint sets for the corresponding  $J_i$ .

If this result set is empty, the unification fails immediately, i.e.  $I_0$  is inconsistent with  $K_0$ . Otherwise, we allow the process to continue, breadth first, *only* with those  $\langle J_i, L_i \rangle$  pairs such that the cardinality of  $L_i$  is strictly less than at the previous iteration. The other members are left unchanged in the final result, where they are interpreted as *provisional* solutions pending arrival of further information, for example at the next step in a derivation.

### 4.2 Decidability

It is evident that, when all steps in a derivation have been completed, the process described above will in general yield a set of information/constraint pairs

$$\{\langle I_1, K_1 \rangle, \dots, \langle I_n, K_n \rangle\}$$

where some solutions are still incomplete—i.e., some of the  $K_i$  are not empty. In very many circumstances it may well be legitimate to take no further action—for example where the output from a linguistic processor will be passed to some other device for further treatment, or where one solution is adequate and at least one of the  $K_i$  is empty. Generally, however, the result set will have to be processed further.

The obvious move, of relaxing the requirement on immediate local convergence and allowing the iteration to proceed without bound, is of course not guaranteed to converge at all in pathological cases. Even so, if there exist some finite number of complete solutions our depth first strategy is guaranteed to find them eventually. If even this expedient fails, or is unacceptable for some reason, the user is allowed to change the environment dynamically so as to set an arbitrary depth bound on the number of final divergent iterations. In these latter cases, the result is presented in the form of a feature structure annotated with details of any constraints which are still unresolved.

### 4.3 Discussion

Designers of unification grammar formalisms have tended to avoid including constructs with the power of relational abstraction, presumably through concern about issues of completeness and decidability. We feel that this is an unfortunate decision in view of the tremendous increase in expressiveness which these constructs can give. (Incidentally, they can be introduced, as in UD, without compromising declarativeness and monotonicity, which are arguably, from a practical point of view, more important considerations.) On a more pragmatic note, UD has been run without observable error on evolving descriptions of substantial subsets of French and German, and it has been rarely necessary to intervene on the depth bound, which defaults to zero.

In practice, users seem to need the extra power very sparingly, perhaps in one or two abstractions in their entire description, but then it seems to be crucially important to the clarity and elegance of the whole descriptive structure (list appending operations, as in HPSG, for example, may be a typical case).

### 4.4 Other extensions

Once we have a mechanism for ‘lazy’ unification, it becomes natural to use the same apparatus to implement a variety of features which improve the habitability and expressiveness of the system as a whole. Most obviously we can exploit the same framework of local convergence or suspension to support efficient hand-coded versions of some basic primitives like list concatenation and non-deterministic extraction of elements from arbitrary list positions. This has been done to advantage in our case, for example, to facilitate importation of useful ideas from, *inter alia* HPSG and JPSG (Gunji, 1987). We have also implemented a fully generalised disjunction (as opposed to the atomic value disjunction described in section 2) using the same lazy strategy to avoid exploding alternatives unnecessarily. Similarly, it was quite simple to add a treatment of underspecified pathnames to allow simulation of some recent ideas from LFG (Kaplan, Maxwell and Zaenen, 1987).

## 4.5 Current state of the system

The system has now been in a stable state for some years, and supports substantial fragments of German French and Italian. A derivative, ELU, specialised for machine translation applications, has been built at ISSCO, Geneva (see Estival, 1990).

There is also a rich user environment, of which space limitations preclude discussion here, including tracing and debugging tools and a variety of interactive parameterisations for modifying run-time behaviour and performance. The whole package runs on any Unix platform which supports Allegro Common Lisp.

## References

- [1] Bresnan J., ed. *The Mental Representation of Grammatical Relations*, Cambridge, Ma.:MIT Press, 1982.
- [2] D'orre, J. and A. Eisele. "A comprehensive unification-based grammar formalism", DYANA deliverable R3.1.B, Centre for Cognitive Science, University of Edinburgh, Scotland, January 1991.
- [3] Estival, D. "ELU user manual", Technical Report, ISSCO, University of Geneva, 1990.
- [4] Fenstad J-E., P-K. Halvorsen, T. Langholm and J. van Benthem, *Situations, Language and Logic*, Reidel, 1987.
- [5] Gunji T., *Japanese Phrase Structure Grammar*, Reidel, 1987.
- [6] Johnson, R. and C. J. Rupp. "Evaluating complex constraints in linguistic formalisms", In Trost, H., editor, *Feature Formalisms and Linguistic Ambiguity*. Ellis Horwood, Chichester, 1993. To appear.
- [7] Kaplan R., J. Maxwell and A. Zaenen, "Functional Uncertainty", in CSLI Monthly, January 1987.
- [8] Sag I. and C. Pollard, "Head-Driven Phrase Structure Grammar: an Informal Synopsis", CSLI Report no.CSLI-87-79, 1987.
- [9] Rupp, C. J. *Semantic Representation in a Unification Environment*, PhD thesis, University of Manchester, 1990.
- [10] Rupp, C.J., R. Johnson and M. Rosner, "Situation schemata and linguistic representation", in M. Rosner and R. Johnson (eds.), *Computational Linguistics and Formal Semantics*, Cambridge:Cambridge University Press, 1992.
- [11] Shieber S., "The design of a computer language for linguistic information", Proceedings of Coling 84, Stanford, 1984.

# Evaluating English Sentences in a Logical Model\*

Joyce Friedman  
Boston University

Douglas B. Moran  
S.R.I. International

David S. Warren  
S.U.N.Y. Stony Brook

e-mail: *jbf@cs.bu.edu*

## Abstract

A central problem in computational linguistics is to find the meaning of sentences of natural language. We describe a computer program that implements the truth-conditional approach to the meaning of natural language sentences for a fragment of English. The program (1) defines interactively a possible worlds model of the universe, (2) reads in an English sentence and finds its syntactic structures, (3) finds the logical formula corresponding to each syntactic structure, and (4) evaluates the formula in the model. The system implements a logico-linguistic theory based on the work of R. Montague. The main computational contributions are an interactive system for defining a possible worlds model, a way of representing partially specified models, a parsing method for Montague grammars, a reduction algorithm for formulas of the intentional logic, and a program to evaluate arbitrary formulas in a model.

## 1 Problem Statement

Montague grammar provided a new approach to the syntax and semantics of natural language. The basic idea is to take the methods of formal semantics from mathematical logic and apply them to the semantics of natural language. In his paper, "The Proper Treatment of Quantification in Ordinary English" (Montague, 1973), known as *PTQ*, Montague illustrates these ideas in application to a specific fragment of English.

In this paper, we give an overview of our computational studies of the system of *PTQ*. Our purpose was to use computer modelling as a tool in understanding *PTQ* and

---

\* At the time this paper was written the authors were at the University of Michigan. The research was supported in part by National Science Foundation Grants BNS 76-23840 and MCS 76-04297. The paper is a revised version of our "Evaluating English Sentences with a Logical Model", presented to the 7th International Conference on Computational Linguistics, Bergen, Norway, August 1978.

also to explore whether this formal model has potential for use in the computer analysis of language. Thus, the problem that we set ourselves is: given the logico-linguistic system of *PTQ*, determine representations and algorithms for a feasible computational version that can be used to find the meanings of an English sentence. We describe our solution to this problem by giving first a brief overview of *PTQ* and our treatment of it. We then give a component by component discussion of our computational system for *PTQ*, pointing out problems, describing solutions, and giving examples. This paper is an overview; the separate algorithms of the components have been described in more detail elsewhere (Friedman, Moran, and Warren 1978a; Friedman and Warren 1978a, 1979, 1980).

## 2 Montague's *PTQ*

In his paper, "The Proper Treatment of Quantification in Ordinary English" (Montague, 1973), Montague defines a relationship between English sentences and their meanings. More precisely, in *PTQ* Montague does the following:

1. Defines a fragment of English *syntax*, by giving an inductive definition of a set of English sentences with their analysis trees.
2. Defines the syntax of an intensional *logic*, by giving an inductive definition of the set of formulas.
3. Defines a *model* for the intensional logic, and defines the intension and extension of a formula with respect to a model.
4. Gives *translation rules* that define a direct translation of an analysis tree into a formula.
5. Gives *meaning postulates* to restrict the set of models, and definitions for *extensional forms* of words, and examples of their use in simplified formulas.

The *PTQ* system described here is the result of converting these definitions into computational processes. The system is unique in carrying out the full process beginning with parsing English sentences and going through interpretation in a model. We know of two other *PTQ* programs carrying out parts of this enterprise. Theo Janssen's program (Janssen 1976, 1980) began by generating syntactic structures according to *PTQ*, translated them into corresponding formulas, and reduced them to the simplified forms. Bipin Indurkhy's program (Indurkhy 1981) was subsequent to ours, and carried through the process from an input sentence to an extensionalized form.

## 3 The Process Version

In the process version of *PTQ* we impose a directed processing view: to go from an English sentence to its meaning. The program (1) interactively defines a possible worlds model of the universe, (2) reads in an English sentence and finds its syntactic structures,

(3) finds the logical formula corresponding to each syntactic structure, and (4) evaluates the formula in the model. The process version implements Montague's definitions of sentences, logical formulas, and translation between them. It modifies the definitions of model and interpretation. It adds algorithms for parsing, lambda-reduction, semantic-reduction (to be defined), specification of a model, and interpretation of formulas in a model.

To implement the parser, it was necessary to modify Montague's grammar to produce only a finite set of analyses for each sentence by eliminating the redundant uninteresting ones, and to construct an algorithm appropriate for this type of grammar. In the logic, the interesting problems were (1) obtaining a lambda-reduction algorithm and a theorem about its properties, and (2) defining a semantic-reduction algorithm with a proof of its soundness. For the semantic component, the system does not follow Montague in using a standard model. Instead, we use a partially-specified model, defined interactively using named elements and expanded dynamically during use.

## 4 Syntax

Montague defines a grammar inductively. To make a directed process version of Montague grammar, a parser is needed to go from a sentence to its analysis trees.

There are several difficulties to be faced in constructing a parser for Montague grammar. In *PTQ* there are infinitely many parse trees for any sentence; the parser should produce only a finite set, which should contain all the trees that are interestingly different. Also, the grammar of *PTQ*, while basically a context-free grammar, also contains non-context-free substitution rules. For the context-free rules the parser should be efficient, and also should handle left-recursion. The substitution rules are new and complex, and special treatment is required to handle them.

For example, consider the sentence **Every man loves a woman**. The parser obtains seven parses for this sentence, of which the first two are particularly interesting. These are shown in Figure 1 and Figure 2. (Figures in the text are simplified to display particular points.) The first parse uses only context-free rules, and forms the *SENTENCE* by combining the *TERM*, **every man**, with the *VERB-PHRASE*, **love a woman**. The second parse illustrates a substitution rule. The sentence is obtained by a *SUBSTITUTION* which puts the *TERM*, **a woman**, for the *VARIABLE*, **he0**, in the *SENTENCE*, **every man loves him0**. There are infinitely many parses for this sentence in *PTQ*.

The next figures show three ways in which any parse can be made into infinitely many more. First, the rules allow the substitution of one variable for another (Figure 3). Such parses introduce no new meanings and no interesting new structures and thus can be ignored. Two parses which are the same except for different bound variables likewise have the same meaning and only one of such a set is needed (Figure 4). Finally, *PTQ* allows vacuous substitution: a term can be substituted for a variable which does not in fact occur. Such parses can lead to obviously incorrect meanings, so they seem to be simply an error in *PTQ*; for example, with vacuous substitution the sentence **John runs**. has a parse which implies the existence of a unicorn (Figure 5).

The process version of *PTQ* uses an ATN parser which finds only finitely many trees for any sentence (Friedman and Warren 1978). All interesting trees are found, and, in particular, no meanings are lost. For any missing tree there is a tree like it except possibly

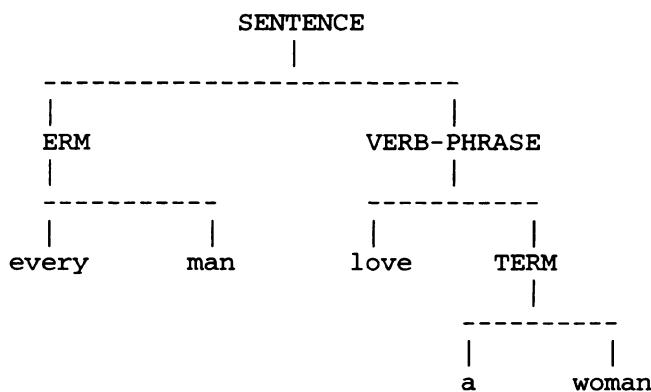


Figure 1: Parse tree for **Every man loves a woman**

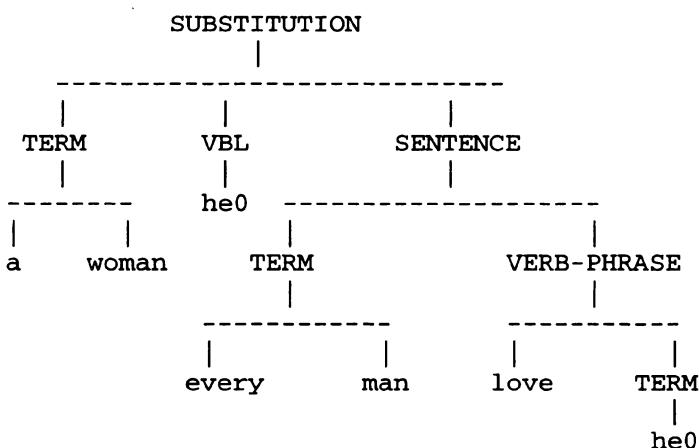


Figure 2: Another parse tree for **Every man loves a woman**

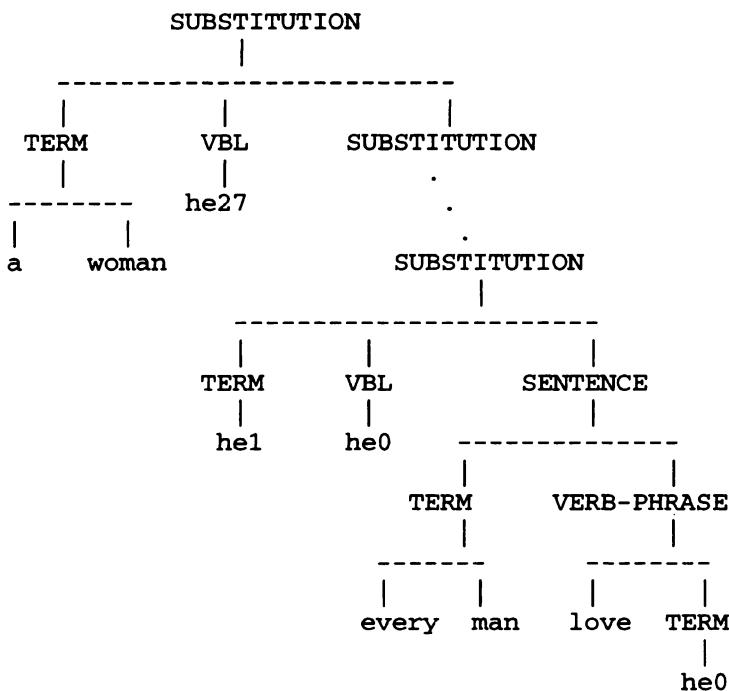


Figure 3: Substitution of variable for variable

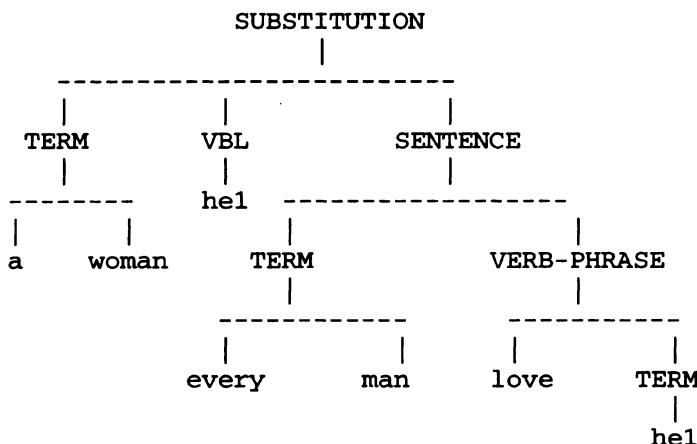


Figure 4: Change of bound variable

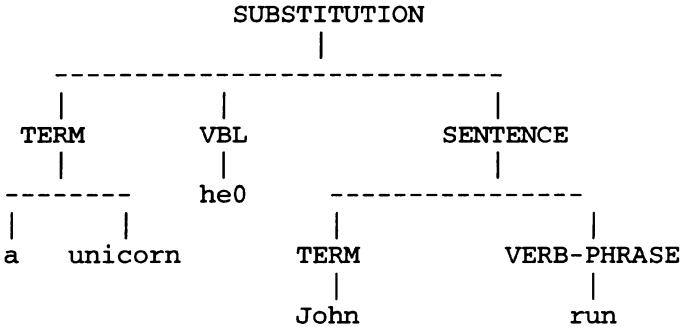


Figure 5: Vacuous substitution

for substitutions. The parser uses an extension of the notion of well-formed substring table to obtain efficiency and to handle left recursion. A special treatment of substitution rules establishes the correct relation between substituted terms and variables.

In this connection we note that the allowable substitutions are not totally arbitrary. Figure 6 shows a “false” parse for the sentence **A woman such-that she loves a man such-that he loves her walks**. This analysis would yield the sentence **A woman such-that she loves a man such-that he loves him0 walks**. This is not the original sentence; it contains the free variable **him0** instead of the pronoun **her**. This example illustrates the need for a constraint on substitution.

Without dwelling on the details of the example, we note that it contains two embedded sentences with associated variables, and that the problem lies in the order of binding the variables. The parser contains a device for avoiding this parse.

(An alternative approach to this problem would be to implement ‘Cooper storage’, introduced in Cooper (1975, 1978) and developed in subsequent papers.)

Figure 7 shows the simplest correct parse for this sentence.

## 5 Translation to Intensional Logic

In *PTQ* Montague gives translation rules which can be implemented in a straight-forward manner. They yield a ‘direct translation’ that is essentially a word for word translation of the sentence. These translations are unreduced expressions of the typed lambda calculus; they are long and complex to interpret and often use higher logical types than are necessary. Our solution to these problems has been to define a lambda-reduction algorithm to carry out syntactic reductions, and to define a \*-reduction algorithm to carry out semantic reductions as justified by the meaning postulates. Although these reduced forms are used in *PTQ*, no algorithm for obtaining them is provided. In Figure 8 we show the results of each of these algorithms for the English sentence **Mary walks**, with the analysis obtained using only context-free rules. The direct translation corresponding to this analysis tree is  $[\lambda P[^V P](^m)](^{\text{walk}})$ . The formula states that the set of properties

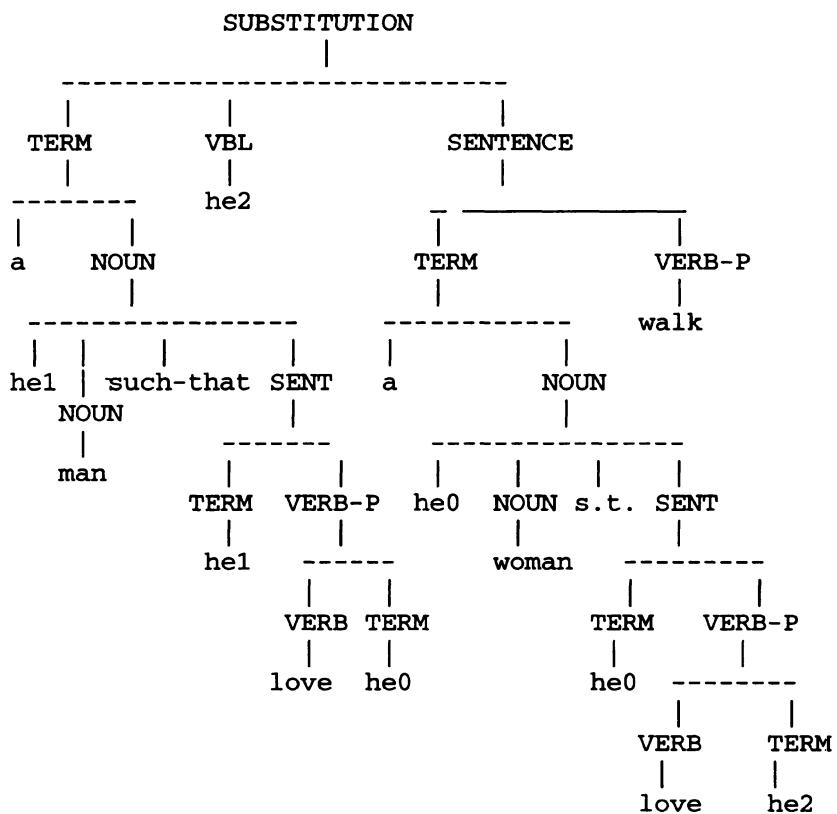


Figure 6: An incorrect parse tree

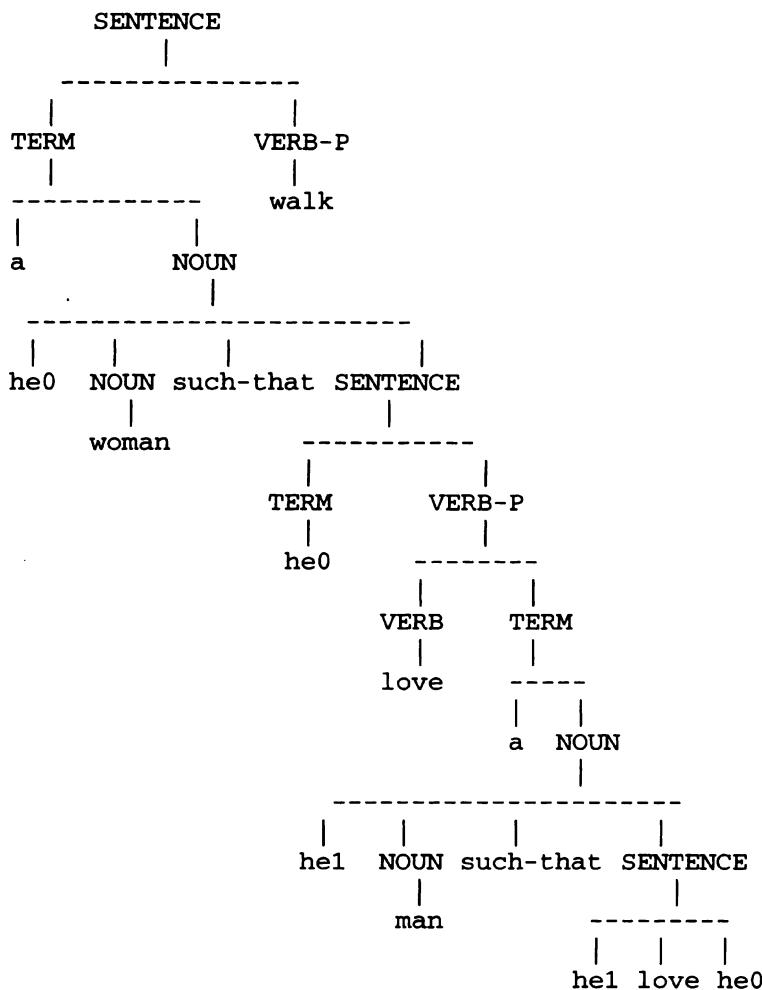
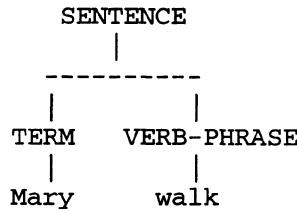


Figure 7: Correct parse

## PARSE TREE



DIRECT TRANSLATION  $[\lambda P[\forall P](^m)](^{\text{walk}'})$

LAMBDA REDUCTION

STEP 1:  
RESULT:  $[\forall (^{\text{walk}'})](^m)$   
 $\text{walk}'(^m)$

\*-INTRODUCTION  $\text{walk}'_*(m)$

Figure 8: Computations for **Mary walks**

corresponding to **Mary** contains the property for **walk'**. The lambda-reduced form of this expression is  $\text{walk}'(^m)$ , which asserts that **walk'** is true of the intension of the constant corresponding to **Mary**. The \*-introduced form is  $\text{walk}'_*(m)$ , which asserts that the extensional verb **walk'\_\*** is true of the entity corresponding to **Mary**. This form allows us to simplify by eliminating intension where there is an equivalent extensional form.

### 5.1 Lambda-normal form theorem

The use of the lambda-reduction algorithm is justified by the following lambda-normal form theorem (Friedman and Warren 1980):

For each formula  $x$  of  $PTQ$ , there is a normal-form formula  $y$ , such that: (1)  $x$  converts to  $y$  by lambda-contraction and  $\wedge^\wedge$ -contraction, (2)  $y$  has no contractible parts, and (3)  $y$  is unique to within change of bound variable.

We note that this theorem is not true in general for the full intensional logic.

### 5.2 Semantic-Reduction or \*-Introduction

In this section we use the convention that  $u$  is a variable of type  $e$  and  $x$  is a variable of type  $< s, e >$ . Montague defines extensional forms for individual words; for example, the extensional form **walk'\_\*** is defined as  $\lambda u \text{ walk}'(^u)$ . He also gives the meaning postulate,  $\exists M \forall x \square [\text{walk}'(x) \leftrightarrow [\forall M](^x)]$ . These yield the theorem,  $\forall x \square [\text{walk}'(x) \leftrightarrow \text{walk}'_*(^x)]$ . To carry out the corresponding transformation, the algorithm first reduces  $x$  to  $^u$  and then introduces **walk'\_\***.

## STANDARD MODEL

$$D_{\langle a,b \rangle} = D_b^{D_a}$$

## GENERALIZED MODEL

$$\emptyset \neq D_{\langle a,b \rangle} \subseteq D_b^{D_a}$$

Closed under lambda-abstraction

## DYNAMIC MODEL

$$D_{\langle a,b \rangle} \subseteq D_b^{D_a}$$

Self-closing under lambda-abstraction

Figure 9: Dynamic models

For example, consider the sentence **John loves Mary**. For each of the seven parses there is a different direct translation. However, the lambda-reduced form of each of these translations is the same:  $\text{love}'(\wedge j, \wedge [\lambda P[\vee P](\wedge m)])$ . The final semantically reduced form is the much simpler  $\text{love}'_*(j, m)$ .

Substitution of the extensionalized form is done only in cases where the result is equivalent under the meaning postulates. For intransitive verbs such as **walk**, the meaning postulates justify substitution of  $\text{walk}'_*(\vee x)$  for  $\text{walk}'(x)$  in all contexts. However, for common nouns, the corresponding substitution is not always justified, so a somewhat complex algorithm was written (see Friedman and Warren, 1979).

The sentence **A unicorn such-that every woman loves it changes** has several distinct logical translations. One of them,  $(\exists x)[(\forall y)[\text{woman}'(y) \rightarrow \text{unicorn}'(x) \wedge [\text{love}'(\wedge \lambda P[\vee P](x))](y)] \wedge \text{change}'(x)]$ , has a reading on which if there are no women, then something changes. For this formula, changing to the extensional form  $\text{unicorn}'_*(\vee x)$  would give an incorrect result. The algorithm does not extensionalize in this case.

## 6 Models for the Intensional Logic

*PTQ* defines a ‘standard model’ and defines interpretation of a formula with respect to a standard model. Our task was to redefine the notion of model so as to make its use computationally feasible. This means that the model must be finitely representable, that its specification must be convenient, and that it must support an appropriate implementation of interpretation in a model. Our solution was to define a class of models that are dynamic, partially-specified, and consist of named elements. The system allows expansion of a model during interpretation of a formula. The model is thus finite, but potentially infinite. Figure 9 shows the relationship between the dynamic models and the usual standard and generalized models of intensional logic. For each type a standard model contains all the possible functions. A generalized model may fail to contain all possible functions, but it is closed under lambda-abstraction. The dynamic models are not closed under lambda-abstraction, but are ‘self-closing’, that is, the system provides

ENGLISH			
	man	walk	John
LOGIC			
	$j \ m \ b$	$\text{man}'_*$	$\text{walk}'_*$
			$\lambda PP(j)$
MODEL			
	JO MA BI	MEN = {JO BI}	WALKERS = {JO MA BI?}

Figure 10: A modest model

for constrained expansion of the model during interpretation of a lambda-expression.

## 6.1 Interactive Specification of a Model

Particular attention was paid to leading the user smoothly through specification of a model. The system directs the specification of a model by asking for entries as they are ready to be defined. We illustrate this here using an extensional model. We emphasize that this simplification is for expository purposes only; the system allows the user to specify either an extensional or an intensional model.

Figure 10 shows a model for a small part of *PTQ*. The formulas to be interpreted in this model contain the English words, **man**, **walk**, and **John**. The corresponding logical formulas contain the constants,  $j$ ,  $m$ , and  $b$  of entity type, and the constants  $\text{man}'_*$  and  $\text{walk}'_*$  of higher type. In the translation to logic **man** becomes  $\text{man}'_*$ , **walk** becomes  $\text{walk}'_*$ , and **John** becomes the formula  $\lambda PP(j)$ . In the model there are three entities, JO, MA, and BI, which are the meanings of the constants  $j$ ,  $m$ , and  $b$ . There is a set MEN consisting just of JO and BI, and a set WALKERS consisting of JO and MA but unspecified for BI.

The interaction by which this model is specified is shown in Figure 11. The program prompts the user for information in an appropriate order. In the figure the user's responses follow the question marks. The program first asks for the set  $D_e$  of entities. The user responds with (JO MA BI). The program then asks for the meaning function F which maps logical constants to elements of the model. It here prompts for the values of F which will be in  $D_e$ , thus it asks for  $F(j)$ ,  $F(m)$ , and  $F(b)$ . The user enters JO, MA, and BI, respectively. The next level of the model is sets of entities. These are functions from  $D_e$  to truth-values. The program asks for a name and the user enters MEN. Then it asks for a value for each element of the domain  $D_e$ . In this case the user specifies that JO and BI are in the set MEN, and MA is not. On a second request for a name, the user enters WALKERS and specifies that JO and MA are in this set and that the membership of BI is not specified. Now the meaning function F can be extended to the constants  $\text{man}'_*$  and  $\text{walk}'_*$ ; the user gives the values MEN and WALKERS.

ENTITIES            D\_e ?            (JO MA BI)

MEANING FUNCTION     F: Logic -> Model

F(j) ?	JO
F(m) ?	MA
F(b) ?	BI

SETS OF ENTITIES,  
functions from D\_e to truth-values

name ?	MEN
value for	
JO ?	1
MA ?	0
BI ?	1
MEN entered	

name ?	WALKERS
value for	
JO ?	1
MA ?	1
BI ?	NIL
WALKERS entered	

MEANING FUNCTION     F: Logic -> Model

F(man*) ?	MEN
F(walk*) ?	WALKERS

Figure 11: Interactive specification of the model

## 6.2 Interpretation in a Model

In interpreting a formula, quantification is restricted to the elements in the named model. This modest model will suffice to illustrate the interpretation of the English sentence **Every man walks**. A corresponding logical formula as determined by the parsing and translation processes is  $(\forall u)[(\text{man}'_*(u) \supset \text{walk}'_*(u))]$ . Figure 12 shows the interpretation of this formula,

and illustrates expansion of the model during interpretation. The quantifier (*forall u*) has as its range the set {JO,MA,BI}. Each of these is taken in turn as the value of *u*. For *u* = JO,  $\text{man}'_*(u)$  is true and  $\text{walk}'_*(u)$  is true, hence the formula is true. For *u* = BI,  $\text{man}'_*(u)$  is true, but the value of  $\text{walk}'_*(u)$  is unknown to the system. Interpretation is interrupted and the program requests that the user enter a value of WALKERS for the element BI. Here the user enters 1 and the system now knows that the set WALKERS is {JO, MA, BI}. Interpretation resumes and the value *true* is returned for the original sentence.

A final more complicated example shows dynamic expansion of the named model during the interpretation of a lambda-function. The logical forms to be interpreted in this model are *j*,  $\text{man}'_*$ ,  $\text{walk}'_*$ ,  $\text{talk}'_*$ , and  $\lambda P P(j)$ . We assume as the initial model the one obtained above in which JO corresponds to *j*, the set MEN = {JO,BI} to  $\text{man}'_*$ , and the set WALKERS = {JO, MA, BI} to  $\text{walk}'_*$ . There is no element in the model corresponding to  $\text{talk}'_*$ . We assume in addition that the model contains one term, TERM1 = {MEN,WALKERS}. We evaluate the (extensional) direct translation of the sentence **John talks**, that is,  $[\lambda P P(j)](\text{talk}'_*)$ . The interpretation process is shown in Figure 13.

First  $[\lambda P P(j)]$  is evaluated. For P = MEN this is true, for P = WALKERS it is true, and the system therefore concludes that its value is TERM1. It now attempts to evaluate  $\text{talk}'_*$ , but since no value is given it asks the user for further specification. It asks first for  $F(\text{talk}'_*)$  and the user gives the new name TALKERS. Since this element is not in the model, the program asks for its specification. The user responds that TALKERS includes JO, is unspecified for MA and does not include BI. Now, since the domain of TERM1 has been expanded, the user must tell the system whether or not TALKERS is one of the properties in TERM1. It is not. Resuming interpretation, the system finds that the lambda expression now evaluates to {MEN,WALKERS,TALKERS}. This set is not a named element of the model; the user is asked to name it so that it can be added. The name TERM2 is given and the evaluation of the formula can then be completed. Since TALKERS is a member of TERM2, the formula is true.

## 7 Conclusion

We have given an overview of a computer program that implements a version of the system of Montague's *PTQ* from an English sentence through to its evaluation in a model. Development of the system required investigation and extension of the syntactic and logical aspects of the system.

In terms of the feasibility of this approach to natural language we can make several observations from this work. The syntactic system was amenable to parsing with only minor adjustments needed. Translation to logic was immediate by a reinterpretation

**INTERPRETATION**

```
(forall u)(implies (man* u) (walk* u))
u = JO
  (implies (man* u) (walk* u))
  (man* u)
    man*
    MEN = {JO,BI}
1
  (walk* u)
  walk*
  WALKERS
1
1
u = MA
  (implies (man* u) (walk* u))
  (man* u)
    man*
    MEN = {JO, BI}
0
1
u = BI
  (implies (man* u) (walk* u))
  (man* u)
    man*
    MEN = {JO, BI}
1
  (walk* u)
  walk*
  WALKERS
= {JO,MA,BI?}
```

**SPECIFICATION**

```
Enter value of WALKERS for BI ?      1
WALKERS = {JO,MA,BI}
```

**RESUME INTERPRETATION**

```
1
1
```

Figure 12: Interpretation of Every man walks

```

INTERPRETATION
  [LAMBDA P P(j)](talk*)
  [LAMBDA P P(j)]
    P = MEN
      P(j)
      1
    P = WALKERS
      P(j)
      1
  TERM1
                talk*
SPECIFICATION
  F(talk*) ?          TALKERS
  TALKERS
    value for
      JO ?           1
      MA ?           NIL
      BI ?           0
  TERM1
    value for
      TALKERS ?       0
RESUME INTERPRETATION
  [LAMBDA P P(j)]
  P = TALKERS
    P(j)
    1
  {MEN, WALKERS, TALKERS}
SPECIFICATION
  NAME THIS ELEMENT ?      TERM2
RESUME INTERPRETATION
  TERM2
                talk*
                TALKERS
  1

```

Figure 13: Interpretation of a formula

of the parse trees. Montague's intensional logic presented some interesting analysis problems. The implementation of the intensional models by finite but expanding representations (dynamic models) allowed an exploration of the notion of interpretation in a model. To apply Montague grammar in a practical natural language system would require much expanded grammars, as well as alternative knowledge representations.

## References

- [1] Cooper, R., *Montague's Semantic Theory and Transformational Syntax*, University of Massachusetts Ph.D. Dissertation, 1975.
- [2] Cooper, R., "A fragment of English with questions and relative clauses", unpublished ms., 1978.
- [3] Friedman, J., D.B. Moran, D. S. Warren. "An interpretation system for Montague Grammar", *American Journal of Computational Linguistics*, microfiche 74, 1978a, 23-96.
- [4] Friedman, J., D.B. Moran, D.S. Warren. "A process version of Montague grammar", Tech. Report N-15, Department of Computer and Communication Sciences, The University of Michigan, Ann Arbor, 1978b.
- [5] Friedman, J., D.B. Moran, D.S. Warren. "Evaluating English sentences with a logical model", *Information Abstracts, 7th International Conference on Computational Linguistics*, Bergen, 1978c, 11pp.
- [6] Friedman, J., D.S. Warren, "A parsing method for Montague Grammars", *Linguistics and Philosophy*, 2, 1978, 347-372.
- [7] Friedman, J., D.S. Warren, "Notes on an intensional logic for English III: extensional forms", Tech. Report N-13, Department of Computer and Communication Sciences, The University of Michigan, Ann Arbor, 1979.
- [8] Friedman, J., D.S. Warren, " $\lambda$ -normal forms in an intensional logic for English", *Studia Logica*, 39(2), 1980, 311-324.
- [9] Indurkyha, Bipin, "Sentence analysis programs based on Montague grammar", M.E.E. Thesis, Philips International Institute of Technological Studies, Eindhoven, 1981.
- [10] Janssen, Theo, "A computer program for Montague grammar: theoretical aspects and proofs for the reduction rules", in J. Groenendijk and M. Stokhof (eds.), *Proc. of the Amsterdam Colloquium on Montague Grammar and Related Topics*. Amsterdam Papers in Formal Grammar I, University of Amsterdam, 1976, 154-176.
- [11] Janssen, T.M.V., "Logical investigations on PTQ arising from programming requirements", *Synthese*, 44, 1980, 361-390.

- [12] Montague, R., “The proper treatment of quantification in ordinary English”, (PTQ), in J. Hintikka, J. Moravcsik, P. Suppes, (eds.), *Approaches to Natural Language*, D. Reidel Publishing Company, Dordrecht, 1973, 221-242; reprinted in R. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague*, Yale University Press, New Haven, 1974, 247-270.

# Recovering Implicit Information\*

Martha S. Palmer, Deborah A. Dahl,  
Rebecca J. Schiffman, Lynette Hirschman,  
Marcia Linebarger, and John Dowding  
R&D, SDC – A Burroughs Company<sup>†</sup>  
*email: mpalmer@linc.cis.upenn.edu*

## Abstract

This paper describes the SDC PUNDIT (Prolog UNDERstands Integrated Text), system for processing natural language messages. PUNDIT, written in Prolog, is a highly modular system consisting of distinct syntactic, semantic and pragmatics components. Each component draws on one or more sets of data, including a lexicon, a broad-coverage grammar of English, semantic verb decompositions, rules mapping between syntactic and semantic constituents, and a domain model.

This paper discusses the communication between the syntactic, semantic and pragmatic modules that is necessary for making implicit linguistic information explicit. The key is letting syntax and semantics recognize missing linguistic entities as implicit entities, so that they can be labeled as such, and reference resolution can be directed to find specific referents for the entities. In this way the task of making implicit linguistic information explicit becomes a subset of the tasks performed by reference resolution. The success of this approach is dependent on marking missing syntactic constituents as elided and missing semantic roles as ESSENTIAL so that reference resolution can know when to look for referents.

## 1 Introduction

This paper describes the SDC PUNDIT<sup>1</sup> system for processing natural language messages. PUNDIT, written in Prolog, is a highly modular system consisting of distinct syntactic, semantic and pragmatics components. Each component draws on one or more sets of data, including a lexicon, a broad-coverage grammar of English, semantic verb decompositions, rules mapping between syntactic and semantic constituents, and a domain model. PUNDIT has been developed cooperatively with the NYU PROTEUS system (Prototype Text Understanding System). These systems are funded by DARPA

\*This work was supported in part by DARPA under contract N00014-85-C-0012, administered by the Office of Naval Research. Approved for public release, distribution unlimited. This paper first appeared in the Proceedings of ACL, 1986, and is reprinted here by permission of the Association for Computational Linguistics. An updated description of this system, including information about different application domains, can be found in Palmer, et al., 1993.

<sup>†</sup>Now known as Unisys Corporation, with whom Deborah Dahl and Marcia Linebarger are still affiliated.

<sup>1</sup>Prolog UNDerstands Integrated Text

as part of the work in natural language understanding for the Strategic Computing Battle Management Program. The PROTEUS/PUNDIT system will map Navy CASREP's (equipment casualty reports) into a database, which is accessed by an expert system to determine overall fleet readiness. PUNDIT has also been applied to the domain of computer maintenance reports, which is discussed here.

The paper focuses on the interaction between the syntactic, semantic and pragmatic modules that is required for the task of making implicit information explicit. We have isolated two types of implicit entities: syntactic entities which are missing syntactic constituents, and semantic entities which are unfilled semantic roles. Some missing entities are optional, and can be ignored. Syntax and semantics have to recognize the OBLIGATORY missing entities and then mark them so that reference resolution knows to find specific referents for those entities, thus making the implicit information explicit. Reference resolution uses two different methods for filling the different types of entities which are also used for general noun phrase reference problems. Implicit syntactic entities, ELIDED CONSTITUENTS, are treated like pronouns, and implicit semantic entities, ESSENTIAL ROLES are treated like definite noun phrases. The pragmatic module as currently implemented consists mainly of a reference resolution component, which is sufficient for the pragmatic issues described in this paper. We are in the process of adding a time module to handle time issues that have arisen during the analysis of the Navy CASREPS.

## 2 The Syntactic Component

The syntactic component has three parts: the grammar, a parsing mechanism to execute the grammar, and a lexicon. The grammar consists of context-free BNF definitions (currently numbering approximately 80) and associated restrictions (approximately 35). The restrictions enforce context-sensitive well-formedness constraints and, in some cases, apply optimization strategies to prevent unnecessary structure-building. Each of these three parts is described further below.

### 2.1 Grammar Coverage

The grammar covers declarative sentences, questions, and sentence fragments. The rules for fragments enable the grammar to parse the "telegraphic" style characteristic of message traffic, such as *disk drive down*, and *has select lock*. The present grammar parses sentence adjuncts, conjunctions, relative clauses, complex complement structures, and a wide variety of nominal structures, including compound nouns, nominalized verbs and embedded clauses.

The syntax produces a detailed surface structure parse of each sentence (where "sentence" is understood to mean the string of words occurring between two periods, whether a full sentence or a fragment). This surface structure is converted into an "intermediate representation" which regularizes the syntactic parse. That is, it eliminates surface structure detail not required for the semantic tasks of enforcing selectional restrictions and developing the final representation of the information content of the sentence. An important part of regularization involves mapping fragment structures onto canonical verb-subject-object patterns, with missing elements flagged. For example,

the two fragment consists of a **tensed verb + object** as in *Replaced spindle motor*. Regularization of this fragment, for example, maps the two syntactic structure into a **verb + subject + object structure**:

*verb(replace),subject(X),object(Y)*

As shown here, **verb** becomes instantiated with the surface verb, e.g., **replace** while the arguments of the **subject** and **object** terms are variables. The semantic information derived from the noun phrase **object spindle motor** becomes associated with **Y**. The absence of a surface subject constituent results in a lack of semantic information pertaining to **X**. This lack causes the semantic and pragmatic components to provide a semantic filler for the missing subject using general pragmatic principles and specific domain knowledge.

## 2.2 Parsing

The grammar uses the Restriction Grammar parsing framework (Hirschman, 1982, Hirschman, 1985), which is a logic grammar with facilities for writing and maintaining large grammars. Restriction Grammar is a descendent of Sager's string grammar (Sager, 1981). It uses a top-down left-to-right parsing strategy, augmented by dynamic rule pruning for efficient parsing (Dowding and Hirschman, 1986). In addition, it uses a meta-grammatical approach to generate definitions for a full range of co-ordinate conjunction structures (Hirschman, 1986).

## 2.3 Lexical Processing

The lexicon contains several thousand entries related to the particular sub-domain of equipment maintenance. It is a modified version of the LSP lexicon with words classified as to part of speech and subcategorized in limited ways (e.g., verbs are subcategorized for the complement types). It also handles multi-word idioms, dates, times and part numbers. The lexicon can be expanded by means of an interactive lexical entry program.

The lexical processor reduces morphological variants to a single root form which is stored with each entry. For example, the form *has* is transformed to the root form *have* in *Has select lock*. In addition, this facility is useful in handling abbreviations: the term *awp* is regularized to the multi-word expression *waiting for part*. This expression in turn is regularized to the root form *wait for part* which takes as a direct object a particular part or part number, as in *is awp 2155-6147*.

Multi-word expressions, which are typical of jargon in specialized domains, are handled as single lexical items. This includes expressions such as *disk drive* or *select lock*, whose meaning within a particular domain is often not readily computed from its component parts. Handling such frozen expressions as "idioms" reduces parse times and number of ambiguities.

Another feature of the lexical processing is the ease with which special forms (such as part numbers or dates) can be handled. A special "forms grammar", written as a definite clause grammar (Pereira, 1980) can parse part numbers, as in *awaiting part 2155-6147*, or complex date and time expressions, as in *disk drive up at 11/17-1236*. During parsing, the forms grammar performs a well-formedness check on these expressions and assigns them their appropriate lexical category.

### 3 Semantics

There are two separate components that perform semantic analysis, NOUN PHRASE SEMANTICS and CLAUSE SEMANTICS. They are each called after parsing the relevant syntactic structure to text semantic well-formedness while producing partial semantic representations. Clause semantics is based on Inference Driven Semantic Analysis (Palmer, 1985) which decomposes verb into component meanings and fills their semantic roles with syntactic constituents. A KNOWLEDGE BASE, the formalization of each domain into logical terms, SEMANTIC PREDICATES, is essential for the effective application of Inference Driven Semantic Analysis, and for the final production of a text representation. The result of the semantic analysis is a set of PARTIALLY instantiated semantic predicates which is similar to a frame representation. To produce this representation, the semantic components share access to a knowledge base, the DOMAIN MODEL, that contains generic descriptions of the domain elements corresponding to the lexical entries. The model includes a detailed representation of the types of assemblies that these elements can occur in. The semantic components are designed to work independently of the particular model, and rely on an interface to ensure a well-defined interaction with the domain model. The domain model, noun phrase semantics and clause semantics are all explained in more detail in the following three subsections.

#### 3.1 Domain Model

The domain currently being modelled by SDC is the Maintenance Report domain. The texts being analyzed are actual maintenance reports as they are called into the Burroughs Telephone Tracking System by the field engineers and typed in by the telephone operator. These reports give information about the customer who has the problem, specific symptoms of the problem, any actions taken by the field engineer to try and correct the problem, and success or failure of such actions. The goal of the text analysis is to automatically generate a data base of maintenance information that can be used to correlate customers to problems, problem types to machines, and so on.

The first step in building a domain model for maintenance reports is to build a semantic net-like representation of the type of machine involved. The machine in the example text given below is the B4700. The possible parts of a B4700 and the associated properties of these parts can be represented by an **isa** hierarchy and a **haspart** hierarchy. These hierarchies are built using four basic predicates: **system**, **isa**, **hasprop**, **haspart**. For example the system itself is indicated by **system(b4700)**. The **isa** predicate associates TYPES with components, such as **isa(spindle^motor,motor)**. Properties are associated with components using the **hasprop** relationship, are inherited by anything of the same type. The main components of the system: **cpu**, **power\_supply**, **disk**, **printer**, **peripherals**, etc., are indicated by **haspart** relations, such as **haspart(b4700,cpu)**, **haspart(b4700,power\_supply)**, **haspart(b4700,disk)**, etc. These parts are themselves divided into sub-parts which are also indicated by **haspart** relations, such as **haspart(power\_supply,converter)**.

This method of representation results in a general description of a computer system. Specific machines represent INSTANCES of this general representation. When a particular report is being processed, **id** relations are created by noun phrase semantics to

associate the specific computer parts being mentioned with the part descriptions from the general machine representation. So a particular B4700 would be indicated by predicates such as these: **id(b4700,system1)**, **id(cpu,cpu1)**, **id(power\_supply,power\_supply1)** etc.

### 3.2 Noun Phrase Semantics

Noun phrase semantics is called by the parser during the parse of a sentence, after each noun phrase has been parsed. It relies heavily on the domain model for both determining semantic well-formedness and building partial semantic representations of the noun phrases. For example, in the sentence, *field engineer replaced disk drive at 11/2/0800*, the phrase *disk drive at 11/2/0800* is a syntactically acceptable noun phrase (as in *participants at the meeting*). However, it is not semantically acceptable in that *at 11/20/800* is intended to designate the time of the replacement, not a property of the disk drive. Noun phrase semantics will inform the parser that the noun phrase is not semantically acceptable, and the parser can then look for another parse. In order for this capability to be fully utilized, however, an extensive set of domain-specific rules about semantic acceptability is required. At present we have only the minimal set used for the development of the basic mechanism. For example, in the case described here, *at 11/2/0800* is excluded as a modifier for *disk drive* by a rule that permits only the name of a location as the object of *at* in a prepositional phrase modifying a noun phrase.

The second function of noun phrase semantics is to create a semantic representation of the noun phrase, which will later be operated on by reference resolution. For example, the semantics for *the bad disk drive* would be represented by the following Prolog clauses.

```
[(id(disk^drive,X),  
  bad(X),  
  def(X),           that is X was referred to with a full,  
  full_npe(X) ]    definite noun phrase,  
                   rather than a pronoun or indefinite noun phrase.
```

### 3.3 Clause semantics

In order to produce the correct predicates and the correct instantiations, the verb is first decomposed into a semantic predicate representation appropriate for the domain. The arguments to the predicates constitute the SEMANTIC ROLES of the verb, which are similar to cases. There are domain specific criteria for selecting a range of semantic roles. In this domain the semantic roles include: **agent**, **instrument**, **theme**, **object1**, **object2**, **symptom** and **mod**. Semantic roles can be filled either by a syntactic constituent supplied by a mapping rule or by reference resolution, requiring close cooperation between semantics and reference resolution. Certain semantic roles are categorized as ESSENTIAL, so that pragmatics knows that they need to be filled if there is no syntactic constituent available. The default categorization is NON-ESSENTIAL, which does not require that the role be filled. Other semantic roles are categorized as NON-SPECIFIC or SPECIFIC depending on whether or not the verb requires a specific referent for that semantic role (see Section 4). The example given in Section 5 illustrates the use of both a non-specific semantic role and an essential semantic role. This section explains

the decompositions of the verbs relevant to the example, and identifies the important semantic roles.

The decomposition of *have* is very domain specific:

```
have(time(Per)) ←  
    symptom(object1(O1),symptom(S),time(Per))
```

It indicates that a particular **symptom** is associated with a particular **object**, as in *the disk drive has select lock*. The **object1** semantic role would be filled by the disk drive, the subject of the clause, and the **symptom** semantic role would be filled by *select lock*, the object of the clause. The **time(Per)** is always passed around, and is occasionally filled by a time adjunct, as in *the disk drive had select lock at 0800*.

In addition to the mapping rules that are used to associate syntactic constituents with semantic roles, there are selection restrictions associated with each semantic role. The selection restrictions for *have* test whether or not the filler of the **object1** role is allowed to have the type of symptom that fills the **symptom** role. For example, only disk drives have select locks.

### 3.3.1 Mapping Rules

The decomposition of *replace* is also a very domain specific decomposition that indicates that an **agent** can use an **instrument** to exchange two **objects**.

```
replace(time(Per)) ←  
    cause(agent(A),  
        use(instrument(I),  
            exchange(object1(O1),object2(O2),time(Per))))
```

The following mapping rule specifies that the **agent** can be indicated by the subject of the clause.

```
agent(A) ← subject(A) / X
```

The mapping rules make use of intuitions about syntactic cues for indicating semantic roles first embodied in the notion of case (Fillmore, 1968, Palmer, 1981). Some of these cues are quite general, while other cues are very verb-specific. The mapping rules can take advantage of generalities like “SUBJECT to AGENT” syntactic cues while still preserving context sensitivities. This is accomplished by making the application of the mapping rules “situation specific” through the use of PREDICATE ENVIRONMENTS. The previous rule is quite general and can be applied to every **agent** semantic role in this domain. This is indicated by the X on the right hand side of the “/” which refers to the predicate environment of the **agent**, i.e., anything. Other rules, such as “WITH-PP to OBJECT2”, are much less general, and can only apply under a set of specific circumstances. The predicate environments for an **object1** and **object2** are specified more explicitly. An **object1** can be the object of the sentence if it is contained in the semantic decomposition of a verb that includes an **agent** and belongs to the *repair* class of verbs. An **object2** can be indicated by a *with* prepositional phrase if it is contained in the semantic decomposition of a *replace* verb:

```
object1(Part1) ← obj(Part1) / cause(agent(A),Repair_event)
```

```

object2(Part2) ←
  pp(with,Part2) /
  cause(agent(A),use(I,exchange(object1(O1),object2(Part2),T)))

```

### 3.3.2 Selection Restrictions

The selection restriction on an **agent** is that it must be a field engineer, and an **instrument** must be a tool. The selection restrictions on the two objects are more complicated, since they must be machine parts, have the same type, and yet also be distinct objects. In addition, the first object must already be associated with something else in a **haspart** relationship, in other words it must already be included in an existing assembly. The opposite must be true of the second object: it must not already be included in an assembly, so it must not be associated with anything else in a **haspart** relationship.

There is also a pragmatic restriction associated with both objects that has not been associated with any of the semantic roles mentioned previously. Both **object1** and **object2** are essential semantic roles. Whether or not they are mentioned explicitly in the sentence, they must be filled, preferably by an entity that has already been mentioned, but if not that, then entities will be created to fill them (Palmer, 1983). This is accomplished by making an explicit call to reference resolution to find referents for essential semantic roles, in the same way that reference resolution is called to find the referent of a noun phrase. This is not done for non-essential roles, such as the **agent** and the **instrument** in the same verb decomposition. If they are not mentioned, they are simply left unfilled. The **instrument** is rarely mentioned, and the **agent** could easily be left out, as in *The disk drive was replaced at 0800*.<sup>2</sup> In other domains, the **agent** might be classified as obligatory, and then it would have to be filled in.

There is another semantic role that has an important pragmatic restriction on it in this example, the **object2** semantic role in *wait^for^part (awp)*.

```

idiomVerb(wait^for^part,time(Per)) ←
  ordered(object1(O1),object2(O2),time(Per))

```

The semantics of *wait^for^parts* indicates that a particular type of part has been ordered, and is expected to arrive. But it is not a specific entity that might have already been mentioned. It is a more abstract object, which is indicated by restricting it to being non-specific. This tells reference resolution that although a syntactic constituent, preferably the object, can and should fill this semantic role, and must be of type **machine-part**, that reference resolution should not try to find a specific referent for it (see Section ??).

The last verb representation that is needed for the example is the representation of *be*:

```

be(time(Per))←
  attribute(theme(T),mod(M),time(Per))

```

In this domain *be* is used to associate predicate adjectives or nominals with an object, as in *disk drive is up* or *spindle motor is bad*. The representation merely indicates that a **modifier** is associated with a **theme** in an attribute relationship. Noun phrase semantics

---

<sup>2</sup>Note that an elided subject is handled quite differently, as in *replaced disk drive*. Then the missing subject is assumed to fill the **agent** role, and an appropriate referent is found by reference resolution.

will eventually produce the same representation for *the bad spindle motor*, although it does not yet.

## 4 Reference Resolution

Reference resolution is the component which keeps track of references to entities in the discourse. It creates labels for entities when they are first directly referred to, or when their existence is implied by the text, and recognizes subsequent references to them. Reference resolution is called from clause semantics when clause semantics is ready to instantiate a semantic role. It is also called from pragmatic restrictions when they specify a referent whose existence is entailed by the meaning of a verb.

The system currently covers many cases of singular and plural noun phrases, pronouns, *one-* anaphora, nominalizations, and non-specific noun phrases; reference resolution also handles adjectives, prepositional phrases and possessive pronouns modifying noun phrases. Noun phrases with and without determiners are accepted. Dates, part numbers, and proper names are handled as special cases. Not yet handled are compound nouns, quantified noun phrases, conjoined noun phrases, relative clauses, and possessive nouns.

The general reference resolution mechanism is described in detail in (Dahl, 1986). In this paper the focus will be on the interaction between reference resolution and clause semantics. The next two sections will discuss how reference resolution is affected by the different types of semantic roles.

### 4.1 Obligatory Constituents and Essential Semantic Roles

A slot for a syntactically obligatory constituent such as the subject appears in the intermediate representation whether or not a subject is overtly present in the sentence. It is possible to have such a slot because the absence of a subject is a syntactic fact, and is recognized by the parser. Clause semantics calls reference resolution for such an implicit constituent in the same way that it calls reference resolution for explicit constituents. Reference resolution treats elided noun phrases exactly as it treats pronouns, that is by instantiating them to the first member of a list of potential pronominal referents, the **FocusList**.

The general treatment of pronouns resembles that of (Sidner, 1979), although there are some important differences, which are discussed in detail in (Dahl, 1986). The hypothesis that elided noun phrases can be treated in much the same way as pronouns is consistent with previous claims by (Gundel, 1980), and (Kameyama, 1985), that in languages which regularly allow zero-np's, the zero corresponds to the focus. If these claims are correct, it is not surprising that in a sublanguage that allows zero-np's, the zero should also correspond to the focus.

After control returns to clause semantics from reference resolution, semantics checks the selectional restrictions for that referent in that semantic role of that verb. If the selectional restrictions fail, backtracking into reference resolution occurs, and the next candidate on the FocusList is instantiated as the referent. This procedure continues until a referent satisfying the selectional restrictions is found. For example, in *Disk drive is*

*down. Has select lock*, the system instantiates the disk drive, which at this point is the first member of the **FocusList**, as the **object1** of *have*:

```
[event39]
have(time(time1))
symptom(object1([drive10]),
        symptom([lock17]),
        time(time1))
```

Essential roles might also not be expressed in the sentence, but their absence cannot be recognized by the parser, since they can be expressed by syntactically optional constituents. For example, in *the field engineer replaced the motor*, the new replacement motor is not mentioned, although in this domain it is classified as semantically essential. With verbs like *replace*, the type to the replacement, *motor*, in this case, is known because it has to be the same type as the replaced object. Reference resolution for these roles is called by pragmatic rules which apply when there is no overt syntactic constituent to fill a semantic role. Reference resolution treats these referents as if they were full noun phrases without determiners. That is, it searches through the context for a previously mentioned entity of the appropriate type, and if it doesn't find one, it creates a new discourse entity. The motivation for treating these as full noun phrases is simply that there is no reason to expect them to be in focus, as there is for elided noun phrases.

## 4.2 Noun Phrases in Non-Specific Contexts

Indefinite noun phrases in contexts like *the field engineer ordered a disk drive* are generally associated with two readings. In the specific reading the disk drive ordered is a particular disk drive, say, the one sitting on a certain shelf in the warehouse. In the non-specific reading, which is more likely in this sentence, no particular disk drive is meant; any disk drive of the appropriate type will do. Handling noun phrases in these contexts requires careful integration of the interaction between semantics and reference resolution, because semantics know about the verbs that create non-specific contexts, and reference resolution knows what to do with noun phrases in these contexts. For these verbs a constraint is associated with the semantics rule for the semantic role **object2** which states that the filler for the **object2** must be non-specific.<sup>3</sup> This constraint is passed to reference resolution, which represents a non-specific noun phrase as having a variable in the place of the pointer, for example, **id(motor,X)**.

Non-specific semantic roles can be illustrated using the **object2** semantic role in *wait for part (awp)*. The part that is being *awaited* is non-specific, i.e., can be any part of the appropriate type. This tells reference resolution not to find a specific referent, so the referent argument of the **id** relationship is left as an uninstantiated variable. The analysis of *fe is awp spindle motor* would fill the **object1** semantic role with **fe1** from **id(fe,fe1)**, and the **object2** semantic role with **X** from **id(spindle^motor,X)**, as in **ordered(object1(fe1),object2(X))**. If the spindle motor is referred to later on in a relationship where it must become specific, then reference resolution can instantiate the variable with an appropriate referent such as **spindle^motor3** (See Section 5.6).

---

<sup>3</sup>The specific reading is not available at present, since it is considered to be unlikely to occur in this domain.

## 5 Sample Text: A Sentence-by-sentence Analysis

The sample text given below is a slightly emended version of a maintenance report. The parenthetical phrases have been inserted. The following summary of an interactive session with PUNDIT illustrates the mechanisms by which the syntactic, semantic and pragmatic components interact to produce a representation of the text.

1. disk drive (was) down (at) 11/16-2305.
2. (has) select lock.
3. spindle motor is bad.
4. (is) awp spindle motor.
5. (disk drive was) up (at) 11/17-1236.
6. replaced spindle motor.

### 5.1 Sentence 1: Disk drive was down at 11/16-2305.

As explained in ?? above, the noun phrase *disk drive* leads to the creation of an **id** of the form: **id(disk^drive,[drive1])**. Because dates and names generally refer to unique entities rather than to exemplars of a general type, their **ids** do not contain a type argument: **date([11/16-1100]),name([paoli])**.

The interpretation of the first sentence of the report depends on the semantic rules for the predicate *be*. The rules for this predicate specify three semantic roles, an **theme** to whom or which is attributed a **modifier**, and the **time**. After a mapping rule in the semantic component of the system instantiates the **theme** semantic role with the sentence subject, *disk drive*, the reference resolution component attempts to identify this referent. Because *disk drive* is in the first sentence of the discourse, no prior references to this entity can be found. Further, this entity is not presupposed by any prior linguistic expressions. However, in the maintenance domain, when a disk drive is referred to it can be assumed to be part of a B4700 computer system. As the system tries to resolve the reference of the noun phrase *disk drive* by looking for previously mentioned disk drives, it finds that the mention of a disk drive presupposes the existence of a system. Since no system has been referred to, a pointer to a system is created at the same time that a pointer to the disk drive is created.

Both entities are now available for future reference. In like fashion, the propositional content of a complete sentence is also made available for future reference. The entities corresponding to propositions are given event labels; thus **event1** is the pointer to the first proposition. The newly created disk drive, system and event entities now appear in the discourse information in the form of a list along with the date.

```
id(event,[event1])
id(disk^drive,[drive1])
date([11/16-2305])
id(system,[system1])
```

Note however, that only those entities which have been explicitly mentioned appear in the **FocusList**:

**FocusList: [[event1],[drive1],[11/16-2305]].**

The propositional entity appears at the head of the focus list followed by the entities mentioned in full noun phrases.<sup>4</sup>

In addition to the representation of the new event, the pragmatic information about the developing discourse now includes information about part-whole relationships, namely that **drive1** is a part which is contained in **system1**.

### Part-Whole Relationships:

**haspart([system1],[drive1])**

The complete representation of **event1**, appearing in the event list in the form shown below, indicates that at the time given in the prepositional phrase *at 11/16-2305* there is a state of affairs denoted as **event1** in which a particular disk drive, i.e., **drive1**, can be described as *down*.

**[event1]**

**be(time([11/16-2305]))**  
**attribute(theme([drive1]),**  
**mod(down),time([11/16-2305]))**

## 5.2 Sentence 2: Has select lock.

The second sentence of the input text is a sentence fragment and is recognized as such by the parser. Currently, the only type of fragment which can be parsed can have a missing subject but must have a complete verb phrase. Before semantic analysis, the output of the parse contains, among other things, the following constituent list: **[subj([X]),obj([Y])]**. That is, the syntactic component represents the arguments of the verb as variables. The fact that there was no overt subject can be recognized by the absence of semantic information associated with X, as discussed in ???. The semantics for the maintenance domain sublanguage specifies that the thematic role instantiated by the direct object of the verb *to have* must be a symptom of the entity referred to by the subject. Reference resolution treats an empty subject much like a pronominal reference, that is, it proposes the first element in the **FocusList** as a possible referent. The first proposed referent, **event1** is rejected by the semantic selectional constraints associated with the verb *have*, which, for this domain, require the role mapped onto the subject to be classified as a machine part and the role mapped onto the direct object to be classified as a symptom. Since the next item in the **FocusList**, **drive1**, is a machine part, it passes the selectional constraint and becomes matched with the empty subject of *has select lock*. Since no select lock has been mentioned previously, the system creates one. For the sentence as a whole then, two entities are newly created: the select lock (**[lock1]**) and the new propositional event (**[event2]**): **id(event,[event2])**, **id(select-lock,[lock1])**. The following representation is added to the event list, and the **FocusList** and Ids are updated appropriately.<sup>5</sup>

---

<sup>4</sup>The order in which full noun phrase mentions are added to the **FocusList** depends on their syntactic function and linear order. For full noun phrases, direct object mentions precede subject mentions followed by all other mentions given in the order in which they occur in the sentence. See (Dahl, 1986), for details.

<sup>5</sup>This version only deals with explicit mentions of time, so for this sentence the time argument is filled in with a gensym that stands for an unknown time period. The current version of PUNDIT uses verb tense and verb semantics to derive implicit time arguments.

```
[event2]
  have(time(time1))
  symptom(object1([drive1]),
  symptom([lock1]),time(time1))
```

### 5.3 Sentence 3: Spindle motor is bad.

In the third sentence of the sample text, a new entity is mentioned, *motor*. Like *disk drive* from sentence 1, *motor* is a dependent entity. However, the entity it presupposes is not a computer system, but rather, a disk drive. The newly mentioned motor becomes associated with the previously mentioned disk drive.

After processing this sentence, the new entity **motor3** is added to the FocusList along with the new proposition **event3**. Now the discourse information about part-whole relationships contains information about both dependent entities, namely that **motor1** is a part of **drive1** and that **drive1** is a part of **system1**.

```
haspart([drive1],[motor1])
haspart(system1,[drive1])
```

### 5.4 Sentence 4: Is awp spindle motor.

*Awp* is an abbreviation for an idiom specific to this domain, *awaiting part*. It has two semantic roles, one of which maps to the sentence subject. The second maps to the direct object, which in this case is the non-specific spindle motor as explained in Section ???. The selectional restriction that the first semantic role of *awp* be an engineer causes the reference resolution component to create a new engineer entity because no engineer has been mentioned previously. After processing this sentence, the list of available entities has been incremented by three:

```
id(event,[event4])
id(part,[_2317])
id(field^engineer,[engineer1])
```

The new event is represented as follows:

```
[event4]
  idiomVerb(waitFor^part,time(time2))
  wait(object1([engineer1]),
  object2([_2317]),time(time2))
```

### 5.5 Sentence 6: Disk drive was up at 11/17-0800

In the emended version of sentence ?? the *disk drive* is presumed to be the same drive referred to previously, that is, **drive1**. The semantic analysis of sentence 5 is very similar to that of sentence 1. As shown in the following event representation, the predicate expressed by the modifier *up* is attributed to the theme **drive1** at the specified time.

[event 5]

```
be(time([11/17-1236]))  
attribute(theme([drive1]),  
        mod(up),time([11/17-1236]))
```

## 5.6 Sentence 6: Replaced spindle motor.

The sixth sentence is another fragment consisting of a verb phrase with no subject. As before, reference resolution tries to find a referent in the current **FocusList** which is a semantically acceptable subject given the thematic structure of the verb and the domain-specific selectional restrictions associated with them. The thematic structure of the verb *replace* includes an **agent** role to be mapped onto the sentence subject. The only **agent** in the maintenance domain is a field engineer. Reference resolution finds the previously mentioned engineer created for *awp spindle motor*, **[engineer1]**. It does not find an instrument, and since this is not an essential role, this is not a problem. It simply fills it in with another gensym that stands for an unknown filler, **unknown1**.

When looking for the referent of a spindle motor to fill the **object1** role, it first finds the non-specific spindle motor also mentioned in the *awp spindle motor* sentence, and a specific referent is found for it. However, this fails the selection restrictions, since although it is a machine part, it is not already associated with an assembly, so backtracking occurs and the referent instantiation is undone. The next spindle motor on the **FocusList** is the one from *spindle motor is bad*, **([motor1])**. This does pass the selection restrictions since it participates in a **haspart** relationship.

The last semantic role to be filled is the **object2** role. Now there is a restriction saying this role must be filled by a machine part of the same type as **object1**, which is not already included in an assembly, viz., the non-specific spindle motor. Reference resolution finds a new referent for it, which automatically instantiates the variable in the **id** term as well. The representation can be decomposed further into the two semantic predicates **missing** and **included**, which indicate the current status of the parts with respect to any existing assemblies. The **haspart** relationships are updated, with the old **haspart** relationship for **[motor1]** being removed, and a new **haspart** relationship for **[motor3]** being added. The final representation of the text will be passed through a filter so that it can be suitably modified for inclusion in a database.

[event6]

```
replace(time(time3))  
cause(agent([engineer1]),  
      use(instrument(unknown1),  
           exchange(object1([motor1]),  
                  object2([motor2]),  
                  time(time3))))  
included(object2([motor2]),time(time3))  
missing(object1([motor1]),time(time3))
```

### Part-Whole Relationships:

```
haspart([drive1],[motor3])  
haspart([system1],[drive1])
```

## 6 Conclusion

This paper has discussed the communication between syntactic, semantic and pragmatic modules that is necessary for making implicit linguistic information explicit. The key is letting syntax and semantics recognize missing linguistic entities as implicit entities, so that they can be marked as such, and reference resolution can be directed to find specific referents for the entities. Implicit entities may be either empty syntactic constituents in sentence fragments or unfilled semantic roles associated with domain-specific verb decompositions. In this way the task of making implicit information explicit becomes a subset of the tasks performed by reference resolution. The success of this approach is dependent on the use of syntactic and semantic categorizations such as ELLIDED and ESSENTIAL which are meaningful to reference resolution, and which can guide reference resolution's decision making process.

### Acknowledgements

We would like to thank Bonnie Webber for her very helpful suggestions on exemplifying semantics/pragmatics cooperation.

## References

- [1] Dahl, D.A., *Focusing and Reference Resolution in PUNDIT*, Presented at AAAI, Philadelphia, 1986.
- [2] Dowding, J. and L. Hirschman, *Dynamic Translation for Rule Pruning in Restriction Grammar*, submitted to AAAI-86, Philadelphia, 1986.
- [3] Fillmore, C.J., "The Case for Case", in E. Bach and R.T. Harms (eds.), *Universals in Linguistic Theory*, Holt, Rinehart, and Winston, New York, 1968.
- [4] Gundel, J.K., "Zero-NP Anaphora in Russian", *Chicago Linguistic Society Parasession on Pronouns and Anaphora*, 1980.
- [5] L. Hirschman and K. Puder, "Restriction Grammar in Prolog", in M. Van Caneghem (ed.), *Proc. of the First International Logic Programming Conference*, Association pour la Diffusion et le Developpement de Prolog, Marseilles, 1982, 85-90.
- [6] Hirschman, L. and K. Puder, "Restriction Grammar: A Prolog Implementation", in D.H.D. Warren and M. VanCaneghem (eds.), *Logic Programming and its Applications*, 1985.
- [7] Hirschman, L., "Conjunction in Meta-Restriction Grammar", *Journal of Logic Programming*, 1986.
- [8] Kameyama, M., "Zero Anaphora: The Case of Japanese", Ph.D. thesis, Stanford University, 1985.
- [9] Palmer, M. "Inference Driven Semantic Analysis", *Proceedings of the National Conference on Artificial Intelligence (AAAI-83)*, Washington, D.C., 1983.

- [10] Palmer, M.S., A Case for Rule Driven Semantic Processing. *Proceedings of the 19th ACL Conference*, June, 1981.
- [11] Palmer, M.S., “Driving Semantics for a Limited Domain”, Ph.D. thesis, University of Edinburgh, 1985.
- [12] Palmer, M.S., et al., “The KERNEL text understanding system,” *Artificial Intelligence*, 63, 1993, 17-68.
- [13] Pereira, F.C.N. and D.H.D. Warren,“ Definite Clause Grammars for Language Analysis – A Survey of the Formalism and a Comparison with Augmented Transition Networks.” *Artificial Intelligence*, 13, 1980, 231-278.
- [14] Sager, N., *Natural Language Information Processing: A Computer Grammar of English and Its Applications*, Addison-Wesley, Reading, MA. 1981.
- [15] Sidner, C.L., “Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse”, *MIT-AI TR-537*, Cambridge, MA, 1979.

# Flexible Generation: Taking the User into Account

Cécile L. Paris

USC/Information Sciences Institute

and

Information Technology Research Institute

University of Brighton

*e-mail:* paris@isi.edu

Vibhu O. Mittal

Intelligent Systems Lab

Department of Computer Science

University of Pittsburgh

*e-mail:* mittal@cs.pitt.edu

## Abstract

Sophisticated computer systems capable of interacting with people using natural language are becoming increasingly common. These systems need to interact with a wide variety of users in different situations. Typically, these systems have access to large amounts of data and must select from these data the information to present to the user. No single generated text will be adequate across all user types and all situations. Certainly, people plan what they will say or write based in part on their knowledge of the listener or intended reader. Similarly, computer systems that produce language must take their listeners/readers into account in order to be effective. In particular, the user's level of knowledge about the domain of discourse is an important factor in this tailoring, if the text provided is to be both informative and understandable to the user. The text should not contain information that is already known or can be easily inferred, nor should it include facts that the user cannot understand. This paper demonstrates the feasibility of incorporating the user's domain knowledge or *user's expertise*, into a text generation system and addresses the issue of how this factor might affect the content, organization and phrasing of a text. We look at two applications domains: (i) generating descriptions of complex physical objects, and (ii) generating documentation for programming languages. We show how a computer generation system can make use of both stereotypical and individualized user models.

## 1 Need for Tailoring

People speak and write differently according to the situation. For example, a surgeon will describe some aspects of a surgical operation differently depending on whether he or she is in a briefing room, actually carrying out the operation and talking with other doctors and nurses, or discussing it afterwards with friends; two doctors reviewing a patient's problem will employ precise and 'technical' medical terms; on the other hand,

---

When I was training to become an emergency medical technician, the physician in charge stressed the importance of using proper medical terminology. Soon after my graduation, I had to transport a boy with a head wound to the hospital, so I radioed in the description: "Ten-year-old male with ten-centimeter laceration on the left occipital region."

The doctor who had instructed me met us in the emergency room. "What happened, son?" he asked the child. "Did you bop your gourd?"

From (*Reader's Digest*, 1989)

---

Figure 1: Employing different terms in talking to different people.

---

**Medical Dictionary:** (Stedman, 1982)

hypophysis: Glandula pituitaria or basilaris; pituitary or master gland; h. cerebri; an unpaired compound gland suspended from the base of the hypothalamus by a short cordlike extension of the infundibulum, the hypophysial (or pituitary) stalk.

---

**Ordinary dictionary** (Webster, 1979):

hypophysis: the pituitary gland.

---

Figure 2: Dictionary entries for doctors and non-doctors

---

when talking to the patient, they will use a different style of language in order for the patient to understand them. This is illustrated quite strikingly in Figures 1 and 2. In both cases, it is fairly safe to assume that the same text would have been inappropriate for the different contexts shown. The boy (in Figure 1) and the layperson for whom the ordinary dictionary is intended would not have understood the more technical terms because of lack of medical knowledge. On the other hand, although the expert may well understand the more naive text, it will in many cases be too general or vague to convey accurate information. More generally, language varies depending on the person to which it is addressed. Importantly, this variation permeates all levels of linguistic realization, from the content and organization of the text as a whole, as in Figure 2, to lexical and syntactic constructions of individual sentences, as in Figure 1. It is thus clear that for a system to be flexible and generate effective presentations for different types of users in different situations, the system must be able to *dynamically* tailor its output to both the given situation and the user (or a specific user type). Although these remarks are equally applicable to natural language systems as well as multi-media systems, this paper focuses on the former.

To tailor a text to a user, a system must have information about the intended user as well as available resources and mechanisms to affect its behavior. Information about the user is typically stored in a *user model*. The effect of stereotypical user models on texts can be obtained by analyzing the differences between texts intended for different audiences. This analysis enables the specification of resources and mechanisms necessary for the system to generate appropriate texts. These specifications of the user model and the mechanisms can then be validated by implementing a system and

evaluating its output with other texts intended for the various user types. In general, natural language generation offers a good test-bed for user modeling issues: we can study texts (or dialogues) produced by humans, analyze their differences and postulate which aspects of the intended reader account for these differences. By then implementing the resulting theory in a system and varying the user model given as input, we can evaluate the texts produced to test and refine the theory.

A user model can contain a variety of facts about a user, including the user's domain knowledge, the user's goals and plans, and various attributes about the user that might help a system in both its problem solving activity and its generation process. In our work, we have been mainly concerned with characterizing how the user's *domain knowledge* can affect a text and modeling this effect in computational systems. The importance of this specific user characteristic for generation has often been noted, (e.g., Lehnert, 1978; McKeown, 1985), and empirical studies have also shown that the effectiveness of a text can depend on it. For example, Feldman and Klausmeier (1974) found that significant differences in the phrasing of texts for fourth- and eighth- grade students were necessary for maximum transfer.

In this paper, we show that the texts aimed at different audiences can be significantly different in their content, organization, and phrasing. It is thus important for a flexible generation system to have information about the user's level of expertise and exploit it when planning a text. This flexibility does not necessarily imply a high cost and does not require a specific processing model, beyond the explicit representation of the user model and of the strategies employed to generate a text. We show that tailoring can be achieved using a variety of generation techniques (schema-based or plan-based approaches), with user models employing different levels of granularity (stereotypes and individualized). We have studied the effect of the user's level of expertise in various application domains. We present here our results from two domains: the descriptions of complex physical objects (such as telephones and radios) and the documentation for a programming language.<sup>1</sup>

## 2 Studying and Representing the Differences in the Texts

We studied the impact of the user's level of expertise on the content and organization of the texts separately from its impact on phrasing. Although this paper concentrates on the former aspect, this section ends by briefly discussing the impact of user modeling on phrasing. To study the effect of the user's expertise on the content and organization of the texts, we obtained texts with the same communicative goal (e.g., describe some object), meant for readers with different levels of expertise, from at least two different sources in each audience category. To study the impact of the user's knowledge on phrasing, we collected texts in which the same information was written for different audiences. In all studies, we examined about twenty to thirty texts from each source.

---

<sup>1</sup>Details about the analyses, results and implementations for these two studies can be found in Paris (1988; 1993) and Mittal (1993).

---

Typical household incandescent lamps (general service) are constructed with the following parts: a coiled tungsten filament, a glass bulb to keep air out and inert gases in, and a base that serves as a holding device and connects the filament to the electric supply. These three parts vary in size and shapes with each different class of lamps, such as general service, reflector, showcase, street-lighting, automobile sealed beam, miniature flash lights and photograph lamps.

*Collier's Encyclopedia* (1962), an adult encyclopedia

---

The type of electric lamp made by Edison is called a filament lamp. A filament lamp lights when a thread inside it heats up to incandescence -- that is when it heats so brightly that it gleams with light.

*The New Book of Knowledge* (1967), a junior encyclopedia

Figure 3: Two descriptions of the filament lamps.

---

## 2.1 Impact on Content and Organization of a Text

In the domain of physical objects, we studied descriptions from various encyclopedias – for both children and adults, (e.g., Collier, 1962; Grolier, 1967), high-school text books, and manuals, (e.g., Chevrolet, 1978; Weissler, 1973). These texts were chosen because they are aimed at audiences at the two ends of the knowledge spectrum: naive and expert. Texts from adult encyclopedias are directed at an audience much more knowledgeable in general than the audience addressed by junior encyclopedias and high school text books. Similarly, the manuals chosen claimed to address either experts or novices. Sample texts are shown in Figure 3.

For documentation examples, our corpus consisted of introductory manuals, tutorials and reference manuals, (e.g., Touretzky, 1984; Steele, 1984), as each tends to be aimed at a different type of audience (introductory manuals for readers without prior knowledge of the programming language, tutorials for either naive or intermediate readers, and reference manuals for experts). Examples of the texts studied are shown in Figure 4 and 5.

We analyzed the texts by decomposing them into different clauses and classifying them as types of rhetorical predicates, (e.g., Shepherd, 1926; Grimes, 1975), as well as performing an analysis in terms of coherence relations, (e.g., Mann & Thompson, 1988), in an attempt to find consistent structures among the texts. In our analysis of texts on programming languages, we also studied the examples, (e.g., the number of examples, the type and amount of information they conveyed, their positions, Mittal, 1993).

Our corpus analysis showed striking, systematic differences between texts aimed at audience with different levels of expertise.

**Differences in the information included** In the domain of physical objects, the analysis showed that the texts fell into two groups: most of the descriptions for expert readers were organized around subparts and their properties, while the descriptions for novices provided functional information, i.e., these descriptions gave an explanation of how the object performed its function (Paris, 1988; 1993).

The texts in the domain of programming languages also showed a clear difference (Mittal & Paris, 1993): the description of programming constructs for advanced

---

A list is recursively defined to be either the empty list or a CONS whose CDR component is a list. The CAR components of the CONSES are called the elements of the list. For each element of the list, there is a CONS. The empty list has no elements at all.  
A list is annotated by writing the elements of the list in order, separated by blank space (space, tab, or return character) and surrounded by parentheses. For example:

```
(a b c)          ; A list of 3 symbols  
(2.0s0 (a 1) #\*) ; A list of 3 things: a short  
                      ; floating point number, another  
                      ; list and a character object
```

Advanced text (Steele, 1984)

---

A list always begins with a left parenthesis. Then come zero or more pieces of data (called the elements of a list) and a right parenthesis. Some examples of lists are:

```
(AARDVARK)  
(RED YELLOW GREEN BLUE)  
(2 3 5 11 19)  
(3 FRENCH FRIES)
```

A list may contain other lists as elements. Given the three lists:

```
(BLUE SKY) (GREEN GRASS) (BROWN EARTH)
```

we can make a list by combining them all with a parentheses.

```
((BLUE SKY) (GREEN GRASS) (BROWN EARTH))
```

Introductory text (Touretzky, 1984)

Figure 4: Definitions from different Lisp texts.

---

---

Used without arguments, **who** lists the login name, terminal name, and login time for each current user. **who** gets this information from the /etc/utmp file.

```
[ ... 16 lines deleted ... ]
```

```
example% who am i  
example!ralph  ttyp0      Apr 27 11:24  
example%  
example% who  
mktg    ttym0    Apr 27 11:11  
gwen    ttyp0    Apr 27 11:25  
ralph   ttyp1    Apr 27 11:30  
example%
```

Text for advanced users from (*Unix Manuals*, 1986).

Figure 5: Sample description from a UNIX manual.

---

users were detailed and complete, including the internal structure of the construct, whereas the descriptions for novice users were syntax oriented. There were differences in the type of examples provided, their number and their complexity as well. Texts from introductory manuals included many simple examples, both positive and negative. On the other hand, texts from reference manuals tended to include only a few complex examples. Reference manuals did not include negative examples, but instead provided anomalous examples,<sup>2</sup> while these were always absent from introductory texts.

**Differences in the organization** As a result of the differences in content, there were of course differences in the organization of texts as well. For object descriptions, one could identify two text organizations corresponding to the two types of information included (functional and structural). But this was not the only difference in textual organization. There were also differences in organization unrelated to the differences in content selection. For example, in our corpus from programming languages, there were clear differences in the respective positions of the examples and the text surrounding them: in the texts for novices, examples were interspersed in the surrounding description, appearing as soon as the feature they illustrated was mentioned. On the other hand, the texts for more advanced users presented their examples grouped at the end of the description (Mittal & Paris, 1993).

These differences are illustrated in the sample texts given above: In Figure 3, the description from the adult encyclopedia provides the parts and attributes of the lamp, while the other description explains how it works. In Figure 4, the reference manual provides a complete description of a list, and only has two examples, given at the end. The second example is complex and illustrates several features: a list can contain several elements; elements can be of various types (e.g., numbers, lists, etc.). In contrast, the description from the introductory book only describes the syntactic feature of a list and provides simple examples as it describes each feature.

## 2.2 Exploiting User Model Constraints to Choose Content and Organization

Having identified these differences, we formalized them in terms of discourse strategies and the user model constraints on their appropriateness. We employed different techniques in the two domains, illustrating how tailoring does not require a specific generation model or a particular representation for the user model.

### 2.2.1 Employing schemas and individual user models

**Representing the discourse strategies** In the domain of physical objects, we employed a simple approach to generating multi-sentential texts: the schema-based approach (McKeown, 1985), using augmented-transition networks (ATN) (Woods, 1973). This approach and formalism were chosen as they provided us with a simple, elegant and efficient way to implement our strategies and were sufficient for our needs. We

---

<sup>2</sup>Anomalous examples represent irregular or exceptional cases. These are instances of the concept being described, but are not covered by the description.

represented the two distinct strategies found in our corpus analysis (corresponding to the different types of texts) in terms of two distinct schemas: (i) the *constituency schema*, similar to the one defined in McKeown's work and consisting of rhetorical predicates, was used to capture texts presenting mostly structural information; (ii) the *process trace*, a schema consisting of directives on tracing the knowledge base, was employed to provide a functional account of an object to represent the functionally-oriented texts (Paris, 1993).

With a schema-based approach, a generation system constructs a text by traversing the ATN, retrieving information from the knowledge base as the arcs of the ATN are chosen. Each time an arc is taken, a proposition corresponding to a rhetorical predicate or a directive is obtained – e.g., (*Identification telephone (device)*) corresponds to applying the predicate *identification* to the object *telephone*, retrieving *device* from the knowledge base). Each such proposition corresponds roughly to a sentence in the text to be generated – e.g., the proposition above would correspond to “A telephone is a device”. After traversing the whole schema, a conceptual representation of the text is obtained. It is passed through an interface which produces the input appropriate for the sentence generator.

When a schema-based system contains several schemas appropriate for the same discourse goal (which is usually the case) it also needs a selection strategy. This is where the constraints on the user model are given. A test on the user model is thus added to other tests that may be applicable, without adding much cost to the process. Note that this can be done either procedurally or declaratively.

**Representing the user model** Had we wanted our system to generate only two types of text, we could have employed a stereotype-approach to the representation of the user model (Rich, 1979), simply indicating whether the user was a novice or an expert user. But we wanted our system to have more flexibility and be able to tailor to users who were not necessarily naive or expert. We noticed in our analysis that texts intended for audiences with knowledge in some areas and not others could be seen as a combination of the two strategies mentioned above (functional and structural). To provide description for users with levels of expertise falling between the extremes of novice and expert, then, our system could be made to combine the strategies. To control the combinations, we also needed to represent users with such intermediate levels of expertise. Defining intermediate stereotypes, however, as in (Chin, 1989), did not appear useful in this domain; it is hard to partition the knowledge base into specific categories and claim that knowledge about a one object type implies greater expertise than knowledge about another, different object type. (For example, there is little reason to believe that knowing about microphones indicates more expertise about the domain as a whole than knowing about telescopes.) As a result, we decided to employ individualized models, indicating explicitly which concepts were known to the user. These models were then exploited by the generation system which was capable of switching strategies depending on whether the user had knowledge about a concept or not. This allowed our system to generate texts for a wide-variety of users spanning a continuum from novice to expert. Sample generated texts (for a naive and an expert user) are shown in Figure 6.

---

A telephone is a device that transmits soundwaves. Because a person speaks into the transmitter of a telephone, a varying current is produced. Then, the current flows through the receiver. This causes soundwaves to be reproduced.

Telephone: Description for a naive user

---

A telephone is a device that transmits soundwaves. The telephone has a housing that has various shapes and colors, a transmitter that changes soundwaves into current, a curly-shaped cord, a line, a receiver to change current into soundwaves and a dialing-mechanism. The transmitter is a microphone with a small disc-shaped thin metal diaphragm. The receiver is a loudspeaker with a small aluminum diaphragm. The housing contains the transmitter and the receiver. The housing is connected to the dialing-mechanism by the cord. The line connects the dialing-mechanism to the wall.

Telephone: Description for an expert user

---

Figure 6: Generated texts in the physical object domain

---

### 2.2.2 Employing a plan-based approach and stereotypes

**Representing the discourse strategies:** In the domain of software documentation, we employed a planning approach instead of the schema-based one, as our implementation was to be part of a larger system which required the flexibility that a planning approach provides (Moore & Paris, 1993). The discourse strategies identified for describing programming constructs were represented as fined-grained plans which captured the intentions behind each portion of the generated text.

In this approach, the generation system constructs text by explicitly reasoning about the communicative goal to be achieved, as well as about the manner in which goals relate to each other rhetorically to form a coherent text.<sup>3</sup> Given a top level communicative goal, such as (KNOW-ABOUT HEARER (CONCEPT LIST)), the system finds plans capable of achieving this goal. Plans typically post further sub-goals to be satisfied. These are expanded, and planning continues to the level of primitive speech acts, e.g., INFORM, which can be directly realized in language without further planning. In this system, the result of the planning process is a discourse tree, where the nodes represent goals at various levels of abstraction, with the root being the initial goal, and the leaves representing primitive realization statements, such as (INFORM ...) statements. The discourse tree also includes coherence relations, which indicate how the various portions of text resulting from the discourse tree will be related rhetorically. This tree is passed to an interface which, as in the previous implementation, converts it into a set of inputs suitable for a sentence generator.

The plan operators employed in this approach can be seen as small schemas which describe how to achieve a goal; like the schemas above, they embody the discourse strategies found in the texts studied. They include conditions for their applicability, which can refer to the system knowledge base, the user model, or the context (the text plan tree under construction and the dialogue history). The constraints on which strategy is applicable for which user model are thus encoded as part of the constraints. A (simplified) plan operator is shown in Figure 7.

---

<sup>3</sup>See Moore & Paris (1993) for details.

---

```
(define-text-plan-operator
  :name introductory-description
  :effect (KNOW-ABOUT hearer (object ?object) (ftr ?ftr))
  :constraints
    (get-user-type ?user-type ?object) ; get stereotypical User Model
    (intro-user-type? ?user-type) ; introductory user?
    (get-example ?example ?object ?ftr) ; can find example?
    (not (in-current-text-plan ; has ?FTR been described?
      (know-about hearer (syntax (object ?object) (ftr ?ftr))))))
    ;; present the syntactic description and an example
  :nucleus (KNOW-ABOUT hearer (syntax (object ?object) (ftr ?ftr)))
  :satellites (ELABORATE (object ?object ftr ?ftr) (example ?example))
```

Figure 7: Simplified plan operator

---

**Representing the user model** We decided that generating documentation for the three user types, novice, intermediate and expert, was sufficient as a start. We thus simply represented our user model in terms of these three stereotypes. The texts for these different audiences exhibited enough differences to warrant such a representation. Thus, the user model employed in our system is very simple. The appropriate strategy is chosen through the constraints of the plan operator, as shown in Figure 7. Sample texts generated by this system are shown in Figure 8. We evaluated the system both by comparing the generated texts with the naturally-occurring ones and with a small experiment in which subjects (novice, intermediate and expert) were given descriptions to study and then asked to answer questions about the descriptions. The results showed that there are significant differences that result in comprehension for novices when they are presented with advanced, reference manual style descriptions. Also, both intermediate and advanced users found the introductory descriptions lacking information they deemed necessary. According to our experiment, then, our discourse strategies and the applicability depending on the user model performed well in selecting and organizing the information to be presented.

### 2.2.3 Summary

To summarize, a flexible text generation system should take into account the user's level of expertise, as this can affect the selection of information, as well as the organization of its presentation. This tailoring can be accomplished using either a schema- or plan-based approach. First, one needs to define the different discourse strategies that are appropriate in the application domain, after having identified them through corpus analysis. The strategies can then be represented in the chosen formalism. Each discourse strategy chooses the appropriate information and organization for specific user types. Constraints on the user model can then be applied to select the appropriate strategy at various points in the generation process. Depending on the variations that need to be produced, the user model can be a stereotypical one or a more detailed individualized one.

---

A list consists of a left parenthesis, followed by zero or more data elements, followed by a right parenthesis. For example:

```
(FISHES)
(PLANES ORANGES PIZZAS PIZZAS)
(MONKEYS PLANES)
(1 3)
(MONKEYS 5 2 FISHES)
```

A list can also be made up of other lists. Consider the three lists:

```
(ORANGES ORANGES)
(AARDVARKS ELEPHANTS)
(FISHES APPLES)
```

These can be used to form the list:

```
((ORANGES ORANGES) (AARDVARKS ELEPHANTS)
(FISHES APPLES))
```

---

A list is defined to be either the empty list or a CONS cell whose CDR component is a list. The CAR components of the CONSES are called the elements of the list. For each element of the list, there is a CONS. The empty list NIL has no elements.

A list consists of a left parenthesis, followed by zero or more data elements, followed by a right parenthesis. Data elements can be either symbols, numbers, characters, strings lists, or a mixture thereof. For example:

```
(apples fishes pizza cars) ; A list of symbols
(#\a 5.1s0 (monkey "abc")) ; A list of 3 things: a character
; a floating point number, and
; a string
```

Figure 8: Generated texts in the documentation domain (for naive and expert users).

---

## 2.3 Tailoring the Phrasing of a Text

In the previous section, we showed the impact of the user's level of expertise on the content and organization of a text and explained how a system could plan a text taking into account this user characteristic. As we mentioned in the introduction, tailoring should occur at all levels of linguistic realization, from the content and organization of the text as a whole to lexical and syntactic constructions of individual sentences (i.e., the *phrasing* of a text). We briefly describe here how the user's level of expertise can also affect the phrasing of a text and how a generation system can control its lexico-grammatical resources to appropriately tailor this phrasing. We have studied this issue more specifically in the documentation domain, where a knowledge base system needs to explain to the user the contents of its knowledge base and the syntax employed to write problem solving plans (Bateman & Paris, 1989; 1991).<sup>4</sup>

As a start, we examined the different forms of language that would be required for a system to interact with the three different user types: experts (system developers who want to make sure that the knowledge base is correctly represented and that the system is working properly), intermediate users (end-users who want to follow the system's reasoning and have some knowledge about the domain but not necessarily about computer science), and novices (students who use the system as a tutoring aid and who are naive with respect to the application domain and to computer science).

Given the same propositional content to express, our analysis showed that there were wide differences in how it should be expressed. For system developers, a form as close as possible to the exact system-internal form was needed, using precise, literal and technical language, as in, for instance, the definition of the concept *faulty-system* shown in (a) of Figure 9. For end-users, however, this definition is rather confusing at best; (b) shows a more appropriate text for such users. Finally, (c) presents a text for naive users.

More specifically, we found at least two types of linguistic variations (besides different choices in vocabulary) between the advanced and introductory descriptions:

1. A difference in the salience of various domain concepts: in texts for advanced users, concepts such as *variables* or the *existence of a process* were given high salience (leading to sentences such as "there exists an O such that..."). On the other hand, these were downgraded in the text for novices. For example, the mathematical concept "there exists one variable such that..." was expressed simply by indicating the type of the variable and providing a determiner (leading to sentences such as "there is an input terminal...").
2. A difference in the grammatical features employed. For example, texts for advanced users had very few pronouns (as compared to introductory descriptions) so as to make the language precise.

We implemented the mechanisms necessary to control the generation process to produce these variations given a specification of the user types. They included an algorithm to determine the head-modifier allocation of the various concepts to express

---

<sup>4</sup>This work was done jointly with John Bateman. It is on-going work whose more general aim is to provide a framework within which it is possible to gain systematic control over phrasing.

---

The system is faulty, if there exists a  $O$  in the set of the output terminals of the system such that the expected value of the signal part of  $O$  does not equal the actual value of the signal part of  $O$  and for all  $I$  in the set of the input terminals of the system, the expected value of the signal part of  $I$  equals the actual value of the signal part of  $I$ .

(a) Text generated for system developers

---

The system is faulty, if all of the expected values of its input terminals equal their actual values and the expected value of one of its output terminals does not equal its actual value.

(b) Text generated for end-users

---

The system is faulty, if its inputs are fine and its output is wrong.

(c) Text generated for students

Figure 9: Examples of variations generated from the same propositional input for different user types.

---

based on the user type, and the dynamic construction of a sub-grammar exhibiting the desired grammatical features for the intended user group. The texts shown in Figure 9 were in fact generated by our system. They are all generated from the same underlying representation and are all appropriate responses to the question: “What is a faulty system?” in a domain concerned with digital circuit diagnosis.

### 3 Conclusions

This paper has argued that User Modeling is important in generation because it affects at least three aspects of natural language communication: (*i*) the information content (as in the functional vs. structural descriptions of physical objects; purely syntactic vs. both syntactic and structural descriptions of programming language concepts), (*ii*) the organization of the information in the message (examples occur at very different positions in introductory texts as compared to examples in reference manuals), and (*iii*) the phrasing of the message.

The paper also described how both stereotypical user models (such as the ones for expert and novice categories) as well as individualized user models can be used to maximize the effectiveness of communication. Such user models can be obtained by a variety of means: complete stereotypical models can be obtained from a detailed analysis of naturally-occurring texts and scenarios, (e.g., Rich, 1979; Chin, 1988), while individualized user models can be incrementally obtained from an analysis of on-line interaction with the user, (e.g., Kobsa, 1984; Kass, 1991). The user models can then be incorporated into relevant decision points during the generation process to ensure the replication of relevant effects in the generated texts. We were concerned in this paper with the impact of the user’s level of expertise on a text. There is also a great deal of work on determining the impact of the user’s plans and goals in providing a cooperative response, (e.g., Allen & Perrault, 1980; Carberry, 1990; Goodman, 1992; McKeown, 1988; Van Beek, 1987). Finally, while the paper only discussed the effect of user modeling on text generation, there is evidence to show that user modeling plays an

equally important role in multi-media generation as well.

### Acknowledgments

The work reported in this paper was supported in part by the DARPA contracts N00039-84-C-0165 and DABT63-91-C-0025, the NSF grants IRI-84-51438 and IRI-9003087, and the NASA-Ames grant NCC 2-520. While writing this paper, Dr. Paris was supported in part by the Commission of the European Union Grant LRE-62009 and the EPSRC Grant J19221, and Dr. Mittal by the National Library of Medicine grant R01 LM05299.

## References

- [1] Paul W. Abrahams, Karl Berry, and Kathryn A. Hargreaves. *TEX for the Impatient*. Addison-Wesley Publishing Co., 1990.
- [2] John A. Bateman and Cécile L. Paris. Phrasing a Text in Terms the User Can Understand. In *Proceedings of IJCAI'89*, Detroit, MI, 1989, pp. 1511–1517.
- [3] John A. Bateman and Cécile L. Paris. Constraining the Deployment of Lexicogrammatical Resources During Text Generation: Towards a Computational Instantiation of Register Theory. In Eija Ventola, Editor, *Functional and Systemic Linguistics: Approaches and Uses*, pp. 81–106. Mouton de Gruyter, Berlin, 1991.
- [4] Sandra Carberry. *Plan recognition in Natural Language Dialogue*. MIT Press Cambridge, MA, 1990.
- [5] Chevrolet. *Chevrolet Service Manual*, General Motors Corporation, 1978, Detroit, MI.
- [6] David N. Chin. KNOME: Modeling What the User Knows in UC. In Alfred Kobsa and Wolfgang Wahlster, editors, *User Models in Dialog Systems*. Springer Verlag, Berlin, 1989.
- [7] *Collier's Encyclopedia*, New York, Crowell-Collier Publishing Company 1962.
- [8] Reader's Digest, September 1989. Contributed by Arlene Shovald.
- [9] Katherine Voerwerk Feldman and Herbert J. Klausmeier. The Effects of Two Kinds of Definitions on the Concept Attainment of Fourth- and Eighth-grade Students. *Journal of Educational Research*, 67(5):219–223, January 1974.
- [10] Bradley A. Goodman and Diane J. Litman. On the Interaction Between Plan Recognition and Intelligent Interfaces. *User Modeling and User-Adapted Interaction*, 2(1–2):55–82, 1992.
- [11] J. E. Grimes. *The Thread of Discourse*. Mouton, The Hague, 1975.
- [12] Robert Kass. Building a User Model Implicitly from a Cooperative Advisory Dialog. *User Modeling and User-Adapted Interaction*, 1(3):203–258, 1991.

- [13] *The New Book of Knowledge - The Children's Encyclopedia*, 1967. Grolier Inc., New York.
- [14] Alfred Kobsa. Generating a User Model from WH-Questions in the VIE-LANG System. Technical Report 84-03, Department of Medical Cybernetics, University of Vienna, 1984.
- [15] Wendy G. Lehnert. *The Process of Question Answering*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.
- [16] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *TEXT*, 8(3):243–281, 1988.
- [17] Kathleen R. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge, England, 1985.
- [18] Kathleen R. McKeown. Generating Goal-Oriented Explanations. *International Journal of Expert Systems*, 1(4):377–395, 1988.
- [19] Vibhu O. Mittal. “Generating Descriptions with Integrated Text and Examples.” PhD thesis, University of Southern California, Los Angeles, CA, 1993.
- [20] Vibhu O. Mittal and Cécile L. Paris. Generating Natural Language Descriptions with Examples: Differences between Introductory and Advanced Texts. *Proceedings of the Eleventh National Conference on Artificial Intelligence – AAAI 93*, 1993.
- [21] Vibhu O. Mittal and Cécile L. Paris. The Placement of Examples in Descriptions: Before, Within or After the Text. *Proceedings of First Pacific Association for Computational Linguistics Conference*, Vancouver, Canada, May 1993, pp. 279–287.
- [22] Johanna D. Moore and Cécile L. Paris. Planning Text for Advisory Dialogues: Capturing Intentional, and Rhetorical Information. *Computational Linguistics*, 19(4):651–694, December 1993.
- [23] Cécile L. Paris. Tailoring Object Descriptions to the User’s Level of Expertise. *Computational Linguistics*, 14(3):64–78, September 1988.
- [24] Cécile L. Paris. *User Modeling in Text Generation*. Frances Pinter, 1993.
- [25] Elaine Rich. User Modeling via Stereotypes. *Cognitive Science*, 3:329–354, 1979.
- [26] UNIX Documentation. UNIX User’s Reference Manual 4.3 Berkeley Software Distribution. Computer Systems Research Group, Computer Science Division, University of California, Berkeley, CA, 1986.
- [27] H. R. Shepherd. *The Fine Art of Writing*. Macmillan, New York, 1926.

- [28] Stedman's Medical Dictionary, 24th Edition. Williams and Wilkins, Baltimore, London, Los Angeles, Sydney, 1982.
- [29] Guy L. Steele Jr. *Common Lisp: The Language*. Digital Press, 1984.
- [30] David S. Touretzky. *LISP: A Gentle Introduction to Symbolic Computation*. Harper & Row, New York, 1984.
- [31] Peter van Beek. A Model for Generating Better Explanations. *Proceedings of ACL 1987*, Palo Alto, California, 1987.
- [32] *Webster's New Twentieth Century Dictionary*, Second Edition. New World Dictionaries. New York, 1979.
- [33] A. Weissler and P. Weissler. *A Woman's Guide to Fixing the Car*. Walker and Company, New York, 1973.
- [34] W. Woods. An Experimental Parsing System for Transition Network Grammars. In R. Rustin, editor, *Natural Language Processing*. Algorithmics Press, New York, 1973.

# Stone Soup and the French Room

Yorick Wilks

Department of Computer Science

University of Sheffield

e-mail: *yorick@dcs.sheffield.ac.uk*

## Abstract

The paper argues that the IBM statistical approach to machine translation has done rather better after a few years than many sceptics believed it could. However, it is neither as novel as its proponents suggest nor is it making claims as clear and simple as they would have us believe. The performance of the purely statistical system (and we discuss what that phrase could mean) has not equaled the performance of SYSTRAN. More importantly, the system is now being shifted to a hybrid that incorporates much of the linguistic information that it was initially claimed by IBM would not be needed for MT. Hence, one might infer that its own proponents do not believe "pure" statistics sufficient for MT of a usable quality. In addition to real limits on the statistical method, there are also strong economic limits imposed by their methodology of data gathering. However, the paper concludes that the IBM group have done the field a great service in pushing these methods far further than before, and by reminding everyone of the virtues of empiricism in the field and the need for large scale gathering of data.

## 1 History

Like connectionism, statistically-based machine translation is a theory one was brought up to believe had been firmly locked away in the attic, but here it is back in the living room. Unlike connectionism, it carries no psychological baggage, in that it seeks to explain nothing and cannot be attacked on grounds of its small scale as connectionist work has been. On the contrary that is how it attacks the rest of us.

It is well known that Western Languages are 50% redundant. Experiment shows that if an average person guesses the successive words in a completely unknown sentence he has to be told only half of them. Experiment shows that this also applies to guessing the successive word-ideas in a foreign language. How can this fact be used in machine translation? (King, 1956).

Alas, that early article told us little by way of an answer and contained virtually no experiments or empirical work. Like IBM's approach it was essentially a continuation of the idea underlying Weaver's original memorandum on MT: that foreign languages were a code to be cracked. I display the quotation as a curiosity, to show that the idea itself is not new and was well known to those who laid the foundations of modern representational linguistics and AI.

I personally never believed Chomsky's arguments in 1957 against other theories than his own any more than I did what he was for: his attacks on statistical and behaviorist methods (as on every thing else, like phrase structure grammars) were always in terms of their failure to give explanations, and I will make no use of such arguments here, noting as I say that how much I resent IBM's use of "linguist" to describe everyone and anyone they are against. There is a great difference between linguistic theory in Chomsky's sense, as motivated entirely by the need to explain, and theories, whether linguistic/AI or whatever, as the basis of procedural, application-engineering-orientated accounts of language. The latter stress testability, procedures, coverage, recovery from error, non-standard language, metaphor, textual context, and the interface to general knowledge structures.

Like many in NLP and AI, I was brought up to oppose linguistic methods on exactly the grounds IBM do: their practitioners were uninterested in performance and success at MT in particular. Indeed, the IBM work to be described here has something in common with Chomsky's views, which formed the post-1957 definition of "linguist". It is clear from Chomsky's description of statistical and Skinnerian methods that he was not at all opposed to relevance/pragmatics/semantics-free methods – he advocated them in fact – it was only that, for Chomsky, the statistical methods advocated at the time were too simple a method to do what he wanted to do with transformational grammars. More recent developments in finite state (as in phrase structure) grammars have shown that Chomsky was simply wrong about the empirical coverage of simple mechanisms.

In the same vein he dismissed statistical theories of language on the ground that sentence pairs like:

the.  
I saw a  
triangular whole.

are equally unlikely but utterly different in that only the first is ungrammatical. It will be clear that the IBM approach discussed here is not in the least attacked by such an observation.

### **Is the debate about empiricism? No.**

Anyone working in MT, by whatever method, must care about success, in so far as that is what defines the task. Given that, the published basis of the debate between rationalism and empiricism in MT is silly: we are all empiricists and, to a similar degree, we are all rationalists, in that we prefer certain methodologies to others and will lapse back to others only when our empiricism forces us to. That applies to both sides in this debate, a point I shall return to.

An important note before continuing: when I refer to IBM machine translation I mean only the systems referred to at the end by Brown et al. IBM as a whole supports many approaches to MT, including McCord's (1989) prolog-based symbolic approach, as well as symbolic systems in Germany and Japan.

## **Is the debate about how we evaluate MT? No.**

In the same vein, I shall not, as some colleagues on my side of the argument would like, jump ship on standard evaluation techniques for MT and claim that only very special and sensitive techniques (usually machine-aided techniques to assist the translator) should in future be used to assess our approach.

MT evaluation is, for all its faults, probably in better shape than MT itself, and we should not change the referee when we happen not to like how part of the game is going. Machine-aided translation (MAT) may be fine stuff, but IBM's approach should be competed with head on by those who disagree with it. In any case, IBM's method could in principle provide, just as any other system could, the first draft translation for a translator to improve on line. The only argument against that is that IBM's would be a less useful first draft *if a user wanted to see why certain translation decisions had been taken*. It is a moot point how important that feature is. However, and this is a point Slocum among others has made many times, the evaluation of MT must in the end be economic not scientific. It is a technology and must give added value to a human task. The ALPAC report, it is often forgotten, was about the economics of contemporary MT, not about its scientific status: the report simply said that MT at that time was not competitive, quality for quality, with human translation.

SYSTRAN won that argument later by showing there was a market for the quality it produced at a given cost. We shall return to this point later, but I make it now because it is one that does tell, in the long run, on the side of those who want to emphasize MAT. But for now, and for any coming showdown between statistically and non-statistically based MT – where the latter will probably have to accept SYSTRAN as their champion for the moment, like it or not – we might as well accept existing “quasi-scientific” evaluation criteria, Cloze tests, test sets of sentences, improvement and acceptability judged by monolingual and bilingual judges, etc. None of us in this debate and this research community are competent to settle the economic battle of the future, decisive though it may be.

## **2 Arguments Not to Use Against IBM**

There are other well known arguments that should not be used against IBM, such as that much natural language is mostly metaphorical and that applies to MT as much as any other NLP task and statistical methods cannot handle it. This is a weak but interesting argument: the awful fact is that IBM cannot even consider a category such as metaphorical use. Everything comes out in the wash, as it were, and it either translates or it does not and you cannot ask why. Much of their success rate of sentences translated acceptably is probably of metaphorical uses. There may be some residual use for this argument concerned with very low frequency types of deviance, as there is for very low frequency words themselves, but no one has yet stated this clearly or shown how their symbolic theory in fact gets such uses right (though many of us have theories of that). IBM resolutely deny the need of any such special theory, for *scale* is all that counts for them.

### 3 What is the State of Play Right Now?

Away with rumor and speculation; what is the true *state of play* at the moment? In recent reported but unpublished DARPA-supervised tests the IBM system CANDIDE did well, but significantly worse than SYSTRAN's French-English system over texts on which neither IBM nor SYSTRAN had trained. Moreover, CANDIDE had far higher standard deviations than SYSTRAN, which is to say that SYSTRAN was far more consistent in its quality (just as the control human translators had the lowest standard deviations across differing texts). French-English is not one of SYSTRAN's best systems but this is still a significant result. It may be unpleasant for those in the symbolic camp, who are sure their own system could, or should, do better than SYSTRAN, to have to cling to it in this competition as the flagship of symbolic MT, but there it is. IBM have taken about 4 years to get to this point. French-English SYSTRAN was getting to about IBM's current levels after 3-4 years of work. IBM would reply that they are an MT system factory, and could do the next language much faster. We shall return to this point.

### 4 What is the Distinctive Claim by IBM About How to Do MT?

We need to establish a ground zero on what the IBM system is: their rhetorical claim is (or perhaps was) that they are a pure statistical system, different from their competitors, glorying in the fact that they did not even need French speakers. By analogy with Searle's Chinese Room, one could call theirs a French Room position: MT without a glimmering of understanding or even knowing that French was the language they were working on! There is no space here for a detailed description of IBM's claims (see Brown et al., 1990, 1991a, 1991b). In essence, the method is an adaptation of one that worked well for speech decoding (Jelinek and Mercer, 1980).

The method establishes three components: (a) a trigram model of English sequences; (b) the same for French; (c) a model of quantitative correspondence of the parts of aligned sentences between French and English. The first two are established from very large monolingual corpora in the two languages, of the order of 100 million words, the third from a corpus of *aligned* sentences in a parallel French-English corpus that are translations of each other. All three were provided by a large machine-readable subset of the French-English parallel corpus of Canadian parliamentary proceedings (Hansard). (1) and (2) are valuable independent of the language pair and could be used in other pairings, which is why they now call the model a *transfer* one. A very rough simplification: an English sentence yields likeliest equivalences for word strings (sub-strings of the English input sentence), i.e., French word strings. The trigram model for French re-arranges these into the most likely order, which is the output French sentence. One of their most striking demonstrations is that their trigram model for French (or English) reliably produces (as the likeliest order for the components) the correct ordering of items for a sentence of ten words or less.

What should be emphasized is the enormous amount of pre-computation that this method requires and, even then, a ten word sentence as input requires an additional hour of computation to produce a translation. This figure will undoubtedly reduce with time and hardware expansion but it gives some idea of the computational intensity of IBM's method.

The facts are now quite different. They have taken in whatever linguistics has helped: morphology tables, sense tagging (which is directional and dependent on the properties of French in particular), a transfer architecture with an intermediate representation, plural listings, and an actual or proposed use of bilingual dictionaries. In one sense, the symbolic case has won: they topped out by pure statistics at around 40% of sentences acceptably translated and then added whatever was necessary from a symbolic approach to upgrade the figures. No one can blame them: it is simply that they have no firm position beyond taking what ever will succeed, and who can object to that?

There is then no theoretical debate at all, and their rhetorical points against symbolic MT are in bad faith. It is Stone Soup: the statistics are in the bottom of the pot but all flavor and progress now come from the odd trimmings of our systems that they pop into the pot.

They are, as it were, wholly pragmatic statisticians: less pure than, say, the Gale group (e.g., Gale & Church 1990) at AT&T: this is easily seen by the IBM introduction of notions like the one they call “informants” where a noun phrase of some sort is sought before a particular text item of interest. This is an interpolation of a highly theoretically-loaded notion into a routine that, until then, had treated all text items as mere uninterpreted symbols.

One could make an analogy here with localist versus distributivist sub-symbolic connectionists: the former, but not the latter, will take on all kinds of categories and representations developed by others for their purposes, without feeling any strong need to discuss their status as artifacts, i.e., how they could have been constructed other than by handcrafting.

This also makes it hard to argue with them. So, also, does their unacademic habit of telling you what they’ve done but not publishing it, allegedly because they are (a) advancing so fast, and (b) have suffered ripoffs. One can sympathize with all this but it makes serious debate very hard.

## 5 The Only Issue

There is only one real issue: is there any natural ceiling of success to *PURE* statistical methods? The shift in their position suggests there is. One might expect some success with those methods on several grounds (and therefore not be as surprised as many are at their success):

- There have been substantial technical advances in statistical methods since King’s day and, of course, in fast hardware to execute such functions, and in disk size to store the corpora.
- The redundancy levels of natural languages like English are around 50% over both words and letters. One might expect well-optimized statistical functions to exploit that to about that limit, with translation as much as another NLP task. One could turn this round in a question to the IBM group: how do they explain why they get, say, 40-50% or so of sentences right, rather than 100%? If their answer refers to the well-known redundancy figure above, then the ceiling comes into view immediately.

If, on the other hand, their answer is that they cannot explain anything, or there is no explaining to do or discussions to have, then their task and methodology is a very odd one indeed. Debate and explanation are made impossible and, where that is so, one is normally outside any rational or scientific realm. It is the world of the witch-doctor: Look – I do what I do and notice that (sometimes) it works.

- According to a conjecture I propounded some years ago, with much anecdotal support, *any theory whatever no matter how bizarre will do some MT*. Hence my surprise level is always low.

## 6 Other Reasons for Expecting a Ceiling to Success with Statistics

Other considerations that suggest there is a ceiling to pure statistical methods are as follows:

1. A parallel with statistical information retrieval may be suggestive here: it generally works below the 80% threshold, and the precision/recall tradeoff seems a barrier to greater success by those methods. Yet it is, by general agreement, an easier task than MT and has been systematically worked on for over 35 years, unlike statistical MT whose career has been intermittent. The relationship of MT to IR is rather like that of sentence parsers to sentence recognizers. A key point to note is how rapid the early successes of IR were, and how slow the optimization of those techniques has been since then!
2. A technical issue here is the degree of their reliance on alignment algorithms as a pre-process: in ACL91 they claimed only 80% correct alignments, in which case how could they exceed the ceiling that that suggests?
3. Their model of a single language is a trigram model because moving up to even one item longer (i.e., a quadgram model) would be computationally prohibitive. This alone must impose a strong constraint on how well they can do in the end, since any language has phenomena that connect outside the three item window. This is agreed by all parties. The only issue is how far one can get with the simple trigram-model (and, as we have seen, it gives a basic 40%), and how far can distance phenomena in syntax be finessed by forms of information caching. One can see the effort to extend the window as enormously ingenious, or patching up what is a basically inadequate model when taken alone.

## 7 The Future: Hybrid Approaches

Given the early success of IBM's methods, the most serious and positive question should be what kinds of *hybrid* approach will do best in the future: coming from the symbolic end, plus statistics, or from a statistical base but inducing, or just taking over, whatever symbolic structures help? For this we can only watch and wait, and possibly help a little here and there. However, there are still some subsidiary considerations.

## 7.1 IBM, SYSTRAN, and the Economics of Corpora

In one sense, what IBM have done is partially automate the SYSTRAN construction process: replacing laborious error feedback with statistical surveys and lexicon construction. And all of us, including SYSTRAN itself, could do the same. However, we must always remember how totally tied IBM are to their Hansard text, the Rosetta Stone, one might say, of modern MT. We should remember, too, that their notion of word sense is only and exactly that of correspondences between different languages, a wholly unintuitive one for many people.

The problem IBM have is that few such vast bilingual corpora are available in languages for which MT is needed. If, however, they had to be constructed by hand, then the economics of what IBM has done would change radically. By bad luck, the languages for which such corpora are available are also languages in which SYSTRAN already has done pretty well, so IBM will have to overtake, then widen the gap with, SYSTRAN's performance a bit before they can be taken seriously from an economic point of view. They may be clever enough to make do with less than the current 100 million word corpora per language, but one would naturally expect quality to decline as they did so.

This resource argument could be very important: Leech has always made the point, with his own statistical tagger, that any move to make higher-level structures available to the tagger always ended up requiring much more text than he had expected.

This observation does not accord with IBM's claims, which are rather the reverse, so an important point to watch in future will be whether IBM will be able to obtain adequate bilingual-corpora for the domain-specialized MT that is most in demand (such as airline reservations or bank billings). Hansard has the advantage of being large but is very very general indeed.

## 7.2 Why the AI Argument About MT Still Has force

The basic AI argument for knowledge-based processing does not admit defeat and retreat, it just regroups. It has to accept Bar Hillel's old anti-MT argument (Bar Hillel, 1960) on its own side – i.e., that as he said, good MT must in the end need knowledge representations. One version of this argument is the primitive psychological one: humans do not do translation by exposure to such vast texts, because they simply have not had such exposure, and in the end how people do things will prove important. Note that this argument makes an empirical claim about human exposure to text that might be hard to substantiate. This argument will cut little ice with our opponents, but there may still be a good argument that we do need representations for tasks in NLP related to MT: e.g. we cannot really imagine doing summarization or question answering by purely statistical methods, can we? There is related practical evidence from message extraction: in the MUC competitions (Lehnert & Sundheim, 1991), the systems that have done best have been hybrids of preference and statistics, such as of Grishman and Lehnert, and not pure systems of either type.

There is the related argument that we need access to representations *at some point* to repair errors. This is hard to make precise but fixing errors makes no sense in the pure IBM paradigm; you just provide more data. One does not have to be a hard line

syntactician to have a sense that rules do exist in some linguistic areas and can need fixing.

### 7.3 Hard Problems Do Not Go Away

There remain, too, crucial classes of cases that seem to need symbolic inference: an old, self-serving, one will do such as “The soldiers fired at the women and I saw several fall” (Wilks, 1975).

I simply cannot imagine how any serious statistical method (e.g., not like “pronouns are usually male so make “several” in a gendered translation agree with soldiers”!) can get the translation of “several” into a gendered language right (where we assume it must be the women who fall from general causality). But again, one must beware here, since presumably any phenomenon whatever will have statistically significant appearances and can be covered by some such function if the scale of the corpus is sufficiently large. This is a truism and goes as much for logical relations between sentences as for morphology. It does not follow that that truism leads to tractable statistics or data gathering. If there could be 75,000-word-long Markov chains, and not merely trigrams (which seem the realistic computational limit) the generation of whole novels would be trivial. It is just not practical to have greater-than-three chains but we need to fight the point in principle as well!

Or, consider the following example (due to Sergei Nirenburg):

#### PRIEST IS CHARGED WITH POPE ATTACK (Lisbon, May 14)

A Spanish priest was charged here today with attempting to murder the Pope. *Juan Fernandez Krohn*, aged 32, was arrested after *a man armed with a bayonet* approached the Pope while he was saying prayers at Fatima on Wednesday night.

According to the police, *Fernandez* told the investigators today he trained for the past six months for the assault. He was alleged to have claimed the Pope ‘looked furious’ on hearing *the priest’s* criticism of his handling of the church’s affairs. If found guilty, *the Spaniard* faces a prison sentence of 15-20 years.

(*The Times* 15 May 1982)

The five italicized phrases all refer to the same man, a vital fact for a translator to know since some of those phrases could not be used in any literal manner in another language (e.g. “the Spaniard” could not be translated word-for-word into Spanish or Russian). It is hard to imagine multiple identity of reference like that having *any* determinable statistical basis.

## 8 Is the Pure Statistics Argument What is Being Debated? No

Everything so far refers to the *pure statistics argument*, from which IBM have now effectively backed off. If the argument is then to be about the deployment of hybrid systems and exactly what data to get from the further induction of rules and categories with statistical functions (e.g., what sort of dictionary to use) then there is really no serious argument at all, just a number of ongoing efforts with slightly differing recipes. Less fun, but maybe more progress, and IBM are to be thanked for helping that shift.

### 8.1 IBM as Pioneers of Data Acquisition

I can add a personal note there: when I worked on what I then called Preference Semantics (Wilks, 1975) at McCarthy's Stanford AI Lab, McCarthy always dealt briefly with any attempt to introduce numerical methods into AI – statistical pattern-matching in machine vision was a constant irritation to him – by saying “Where do all those numbers COME from?” I felt a little guilty as Preference Semantics also required at least link counting. One could now say that IBM’s revival of statistical methods has told us exactly where some of these numbers come from! But that certainly does not imply that the rules that express the numbers are therefore useless or superseded.

This touches on a deep metaphysical point: I mentioned above that we may feel word-sense is a non-bilingual matter, and that we feel that there *are* rules that need fixing sometimes, and so on. Clearly, not everyone feels this. But it is our culture of language study that tells us that rules, senses, metaphors, representations etc. are important and that we cannot imagine all that is just a cultural artifact. An analogy here would be Dennett’s recently (1991) restated theory of human consciousness that suggests that all our explanations of our actions, reason, motives, desires etc. as we articulate them may be no more than fluff on the underlying mechanisms that drive us.

IBM’s work induces the same terror in language theorists, AI researchers and linguists alike: all their dearly-held structures may be just fluff, a thing of schoolmen having no contact with the reality of language. Some of us in AI, long ago, had no such trouble imagining most linguistics was fluff, but do not want the same argument turned round on us, that *all* symbolic structures may have the same status.

Another way of looking at this is how much good IBM are doing us all: by showing us, among other things, that we have not spent enough time thinking about how to acquire, in as automatic a manner as possible, the lexicons and rule bases we use. This has been changing lately, even without IBM’s influence, as can be seen from the large-scale lexical extraction movement of recent years. But IBM’s current attempts to recapitulate, as it were, in the ontogeny of their system, much of the phylogeny of the AI species is a real criticism of how some of us have spent the last twenty years.

We have not given enough attention to knowledge acquisition, and now they are doing it for us. I used to argue that Alers and computational linguists should not be seen as the white-coated laboratory assistants of linguistic theorists (as some linguists used to dream of using us). Similarly, we cannot wait for IBMers to do this dirty work for us while we go on theorizing. Their efforts should change how the rest of us proceed from now on.

## 9 Conclusion: Let Us Declare Victory and Carry on Working

Relax, go on taking the medicine. Brown et al.'s retreat to incorporating symbolic structures show the pure statistics hypothesis has failed. All we should be haggling about now is how best to derive the symbolic structures we use, and will go on using, for machine translation.

### Acknowledgments

In acknowledgement of contributions from James Pustejovsky, Bob Ingria, Bran Boguraev, Sergei Nirenburg, Ted Dunning and others in the CRL natural language processing group.

## References

- [1] Bar-Hillel, Y., "The present status of automatic translation of languages", in J. Alt (ed.), *Advances in Computers* 1, Academic Press, New York, 1960.
- [2] Brown, P.F., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, P. Roossin, "A statistical approach to machine translation", in *Computational Linguistics*, 16, 1990, 79-85.
- [3] Brown, P.F., J. Lai, R. Mercer, "Aligning sentences in parallel corpora", in *Proceedings 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 1991, 169-176.
- [4] Chomsky, N., *Syntactic Structures*, Mouton and Co., The Hague, 1957.
- [5] Dennett, D., *Consciousness Explained*, Bradford Books, Cambridge MA, 1991.
- [6] Gale, W., K. Church, "Poor estimates of context are worse than none", in *Proc. 1990 DARPA Speech and Language Meeting*, Hidden Valley, PA, 1990.
- [7] King, G. "Stochastic methods of mechanical translation", in *Mechanical Translation*, 3, 1956.
- [8] Jelinek, F., R. Mercer, "Interpolated estimation of Markov source parameters from sparse data", in *Proceedings of the Workshop on Pattern Recognition in Practice*, North Holland, Amsterdam, The Netherlands, 1980.
- [9] Lehnert, W., B. Sundheim, "A performance evaluation of text analysis technologies", *AI magazine*, 12, 1991.
- [10] McCord, M., "A new version of the machine translation system LMT", *Literary & Linguistic Computing*, 4, 1989.
- [11] Wilks, Y., "A preferential pattern-matching semantics for natural language understanding", *Artificial Intelligence*, 11, 1975.

LINGUISTICA COMPUTAZIONALE  
SPECIAL ISSUES

*Ordenatores y lengua española*, a cura di M. Novella Catarsi, Daniela Ratti, Antonina Laba e Manuela Lassi, pp. 78 Lit. 55,000.

*Research in Natural language Processing in Italy*, A. Capelli (Ed.), pp. 136. Lit. 40,000.

*The possibilities and Limits of the Computer in producing and Publishing Dictionaries*, (esaurito).

*Computers in Literary and Linguistic Research*, Proceedings of the VII International Symposium of the Association for Literary and Linguistic Computing, Pisa, 1982, pp. 302. Lit. 80,000.

*Studies in Honour of Roberto Busa S.J.*, (voll. IV-V/1984-1985), pp. XXXII-362, 1987. Lit. 85,000.

*Computational Lexicology and Lexicography. Special issue dedicated to Bernard Quemada*, vol. I, pp. XLV-374. Lit. 80,000. Vol. II, pp. XVI-466. Lit. 80,000.

*Synopses of American, European and Japanese Projects Presented at the International Projects Day at Caling 1992*, G. B. Varile, A. Zampolli (Eds.), pp. 74. Lit. 40,000.

IMPRESSO NELLE OFFICINE DI AGNANO PISANO DELLA  
GIARDINI EDITORI E STAMPATORI IN PISA



*Giugno 1994*