Computational lexicology: a research program

by ROBERT A. AMSLER
SRI International
Menlo Park, California

ABSTRACT

Computational lexicology may be defined as the application of computers to the study of the lexicon. Taken in its broadest sense, it would be a multidisciplinary field involving the analysis of man-made dictionaries using computers to study their machine-readable text as well as a study of the computational linguistic content and organization of lexicons for use by natural-language processing programs.

Computational lexicology is an emerging field of study for the 1980s being created by converging trends in other disciplines. This paper attempts to outline some of the applications that a knowledge of computational lexicology will facilitate and the possible means of extending such lexical knowledge.
Computational lexicology may be defined as the application of computers to the study of the lexicon. Taken in its broadest sense, it is a multidisciplinary field involving the analysis of dictionaries written by human beings, using computers to study the machine-readable texts, as well as a study of the computational linguistic content and organization of lexicon for use by natural-language-processing programs. Using the term in this broader sense, I shall refer to the computational lexicology task aspects of computational linguistics, artificial intelligence, cognitive science, and information science that are concerned with the lexicon in computational processes.

I see computational lexicology as an emerging field of study for the 1980s being created by converging trends in other disciplines. Two major reasons for the growth of computational lexicology are that computer typesetting is widely available and machine-readable texts of dictionaries are increasingly available. These primary source materials have also appeared at a time when rapidly declining online disk storage costs and increased acceptance of interactive computation have given rise to natural-language-processing systems needing access to sizable quantities of lexical information in real time.

The rising volume of full-text sources, also a consequence of the increasing availability of computer typesetting, has additionally placed a higher premium on the development of computational systems that can access textual information interactively. Full-text database management, advanced computational linguistic processing systems, and other artificial intelligence applications are striving to encompass new domains and larger text volumes to make full use of this increased availability of text and interactive computation. For these reasons, computational lexicology has emerged as a discipline with both a considerable body of data to begin analyzing and a heightened motivation to provide structured lexical information for existing and future natural-language-processing systems.

Computational lexicology also seems to promise to reveal, in greater depth and detail than ever before possible, the structure and consequences of the organization of the lexicon. Interest in the lexicon from disciplines such as philosophy, linguistics, psychology, and anthropology has never been greater. By introducing computation into the study of the human organization of the lexicon in dictionaries, it has become possible to make new observations and theories about the nature of very ancient and perplexing lexical problems. Computational lexicology sheds new light on the deep semantic relationships involved in the process of defining, the nature of our ability to perform lexical disambiguation from context, and the connectivity of the mental lexicon. Our dictionaries, having evolved over a few centuries of effort, are now as much an artifact of our innate language abilities as they are a product illustrating our civilization's view of the world. Study of the lexicon's organization can consequently reveal basic information about the nature and organization of human knowledge and of our civilization.

Text Streams

Whereas in the past computer technology was applied to solving natural-language information-processing problems for selected segments of textual data, many of which were specifically rendered machine-readable by manual keyboarding, there will soon be more machine-readable textual data than can readily be processed by even the most efficient information-processing algorithms. One may soon be able to seek out and obtain machine-readable text in a specific domain by simply contacting a publisher or a text database vendor instead of contemplating hand entry of the data.

This new volume of text will offer us literally infinite "text streams" of information, and each new publication will be added as a tiny rivulet joining the river of new text data. Just as with a river, there will be opportunities to exploit this massive flow. The equivalent of dams can be set up to gather selected segments of textual data, many of which were specifically rendered machine-readable by manual keyboarding, whereas in the past computer technology was applied to solving natural-language information-processing problems for selected segments of textual data, many of which were specifically rendered machine-readable by manual keyboarding, there will soon be more machine-readable textual data than can readily be processed by even the most efficient information-processing algorithms. One may soon be able to seek out and obtain machine-readable text in a specific domain by simply contacting a publisher or a text database vendor instead of contemplating hand entry of the data.

This new volume of text will offer us literally infinite "text streams" of information, and each new publication will be added as a tiny rivulet joining the river of new text data. Just as with a river, there will be opportunities to exploit this massive flow. The equivalent of dams can be set up to gather selected segments of textual data, many of which were specifically rendered machine-readable by manual keyboarding, whereas in the past computer technology was applied to solving natural-language information-processing problems for selected segments of textual data, many of which were specifically rendered machine-readable by manual keyboarding, there will soon be more machine-readable textual data than can readily be processed by even the most efficient information-processing algorithms. One may soon be able to seek out and obtain machine-readable text in a specific domain by simply contacting a publisher or a text database vendor instead of contemplating hand entry of the data.
of words each month. Others will experience periodic flash
ing and only emit text as publications appear weekly,
monthly, quarterly, etc. Among these will be the electronic
journals, which are machine-readable copies of paper jour-
nals produced by publishers as a result of the computer type-
setting process. Finally, some text streams will be completely
aperiodic, such as books and similar documents. Each of these
will appear almost as a tidal wave and not be subjected to
ordinary flow monitoring or processing, but each become a
special processing task.

Text streams can be monitored by various methods. Indicators
of the flow level and rate can be provided. Sampling to
determine the content by subject matter or level of difficulty
or for other purposes can be performed. Text streams can be
filtered to extract different types of information, either to
separate out unwanted text or to separate the constituents of
the stream by category for different dispositions. Varying fil-
ters can provide some destinations of the text with high-level
tables of contents and others with selected extracted sections.
A text stream filter can be an arbitrary program that processes
text information, monitors for certain types of data, and per-
forms other similar tasks.

Problems Facing Computational Linguistic Systems in the
1980s

Computational linguistics has made vast gains over the pre-
vious two decades in parser design and grammar construction.
Several successful systems for parsing database queries in or-
dinary English or directing robot operations and expert prob-
lem solving already exist. Nevertheless, computational lin-
guistics has not yet reached the point at which parsing can be
performed over unrestricted input text. There are many rea-
sons for this.

Although linguistics has entered the age of computation,
the ultimate basis for any sophisticated computational linguis-
tic technique is the computable knowledge accessible in the
system’s lexicon. The lexicon serves not only as the basis for
the recognition vocabulary of any text-processing system, but
as the indexed repository of the vast array of additional syn-
tactic, semantic, and pragmatic information upon which text-
processing algorithms are based. The scope of this informa-
tion for each lexical entry has been steadily increasing in both
volume and complexity over the past decade. From simple
feature-value lists, it has progressed to advanced nonplanar
graph-theoretic representations and lexicons combining com-
putable code with stored information.

All current natural-language systems use lexicon composed
by hand. The number of person-years necessary to describe
the tens of thousands of word senses required for processing
unrestricted text cannot be completed by any individual re-
search project also concerned with building sophisticated soft-
ware to access this lexicon. This has led to a lexical plateau
in computational linguistic systems that limits lexicon to a few
hundred entries in most cases while researchers concentrate
on improving the cleverness of the parsing algorithms and the
grammatical coverage of the systems. What cannot be
achieved by the application of cleverness is the lexical cov-
4age necessary to process text in the expanded domain of
unrestricted English, which is inundating our electronic and
magnetic media.

Without concentrated effort on computational lexicology
and lexicography, the progress of natural-language-processing
systems toward production use in processing unrestricted En-
glish text or the extension of existing restricted-domain sys-
tems to new domains will be extremely slow. The time may be
rapidly approaching when computational linguistics and lexi-
on building should be undertaken as part of separate re-
search tasks, studied independently.

Progress in Computational Lexicology

The previous two sections provide two motivations for
working on computational lexicology: (1) the desirability of
making use of new full-text sources and (2) the problem of
increasing the size and detail of lexical entries for com-
putational linguistic systems to the extent that they can begin
to cope with the increased quantities of text provided by new
full-text sources. In addition to these two motivations, the
successful completion of analyses of the structure of two
machine-readable dictionaries, the Merriam-Webster New
Pocket Dictionary and The Longman Dictionary of Con-
temporary English has promoted interest from computa-
tional linguistics itself in this type of study. Thus, there
are also new methods available for studying machine-readable
dictionaries.

Computational lexicology seeks to understand the mean-
ings of words by accumulating information about their usages.
This is the basic task performed by lexicographers when they
accumulate citations to use as the basis for writing dictionary
entries. Among the potential items of information which can
be accumulated from examination of text are

- Verbs with which a given noun is used
- Nominal compounds in which a word occurs
- Hyphenated forms of other combining forms in which
  words occur
- Adjectives used with nouns
- Morphological variants in which a word form occurs
- Subject domains in which given terms appear
- First occurrences of new word forms
- The identification of defining contexts for new terms in
  free-running text
- The assimilation of information about word forms from
  many separate text sources
- Frequency data on the number of textual occurrences,
  with examples of infrequent phenomena as well as meth-
ods for recognizing such phenomena when they occur

All of this lexical information can be gathered by examining
machine-readable full-text sources, and it constitutes a new
type of information gathering that is of interest to computa-
tional lexicology.

RESEARCH IN THE ACQUISITION OF LEXICAL
KNOWLEDGE

The following program of research in the new area of com-
putational lexicology outlines several areas in which research
could be performed to build an adequate knowledge of the requirements for natural-language-processing-system lexicons. It is directed primarily at ways to use the computer to extract and organize lexical information, which can then be applied to natural-language-processing tasks. Our knowledge of the lexicon and its parameters is extremely deficient.

**Analysis of Machine-Readable Dictionaries As a Source of Lexical Knowledge**

Lexical knowledge derives from the analysis of the usages of words. Dictionaries are one form in which such analysis has traditionally been presented. Dictionaries are based on written textual instances of word usage, called citations, which lexicographers organize into lexical entries that present the basic information about how such a word can be used in textual contexts. Typically, the information presented in an ordinary dictionary is limited to morphology (spellings, hyphenation points), pronunciation, etymology, and syntax (parts of speech, usage notes) with semantic information being given in informal fashion as part of the entry’s definition. Certain dictionaries in the advanced learner’s class include semo-syntactic pattern codes in their lexical entries.

Even in the cases where the semantic information is not codified by the publishers, but appears as part of the text of dictionary definitions, it is possible to apply advanced computational processing to extract much data for potential text system use.²

**Building Lexical Knowledge from Full-Text Sources**

Second in importance as a source of lexical data to direct access to machine-readable dictionary texts is access to the raw data which form the basis for the construction of dictionaries. This task can be approached in three ways.

- First, one can interview human subjects and use a computer to perform bookkeeping, cross-referencing and question-answering to solicit the needed information in an organized and systematic manner.
- Second, one can attempt to extract dictionary definitions from text using fully automatic language analysis and seeking specific definition indicators in texts.
- Third, one can use ordinary text and apply computational linguistic (morphological and syntactic parsing) and information science techniques (cluster analysis, co-occurrence relationships, frequency counts) to gather and present the raw text to human lexicographers or experts for their assimilation and restructuring into formal analyses of terminology.

For the purposes of this paper, the term full-text source will be used to refer to any form of data which represents the full text of some document. This might be a publisher’s photocopying tape containing the text of an article, chapter, book, or any other source. It may contain tables or figures such as would appear in the full text of the published source. Full-text sources will also include materials not intended for publication, such as mail messages; or word processor output, such as business correspondence.

Among the more standard interpretations of full text is the notion of machine-readable bodies of text specifically assembled for the study of language. Numerous efforts have been undertaken to assemble such bodies of text in many of the world’s major languages.

The task of assembling such a text requires careful planning so that the data selected will be useful to computational lexicology. Simply obtaining a multimegabyte set of words by grabbing all the machine-readable sources available and merging them together hardly provides the basis for scientific observation of the nature of the language as a whole. Thus, ten million words of newspaper stories can be less useful than one million words of carefully sampled text taken from a variety of sources. In this regard, perhaps the most carefully prepared body of English is the one-million-word one prepared by Francis and Kucera at Brown University from full-text sources of 1961.⁷ ¹¹ Additionally, a 5-megabyte body of English based upon third- to ninth-grade textbooks was prepared in 1969 by the American Heritage Publishing Company.⁴

**APPLIED COMPUTATIONAL LEXICOLOGY**

Although this paper is intended to support the case for basic research in computational lexicology, it is also worthwhile examining what would be some applications of such research. There is no limit to the number of areas in which computers are applied to processing natural-language text. To some degree all these areas could be affected by developments in computational lexicology. Among the first to be affected might be fields that already have a considerable emphasis on the availability of words (e.g., spelling correction) or have established uses for lexicons (e.g., content analysis systems, mechanical translation systems, and word processing systems). One promising new area that might benefit would be full-text database retrieval.¹⁴

**Spelling Correction**

Spelling correction is one of the most neglected aspects of natural-language system design. It is a stepchild of the sophisticated computational linguistics systems designed within artificial intelligence and often regarded as a simple instance of “bells and whistles” to be added to a natural-language system after the fact. This situation is changing, because studies are beginning to show that of all the areas in which natural-language systems fail, spelling correction (and unknown lexicon) are two of the most frequent causes of failures. In some systems spelling errors and unknown vocabulary may account for nearly 50% of all failures in system response.²¹

Spelling correction as a technique has recently become the topic of intensive, but largely unreported, research in the word processing field. IBM's DisplayWriter system offers a multilingual spelling correction capability, which, though originally thought to be a simple matter of checking lists of correctly spelled words, turned out to be a significant problem.
when languages other than English were assumed to be correctable with the same algorithm as English. This is because spelling correction depends greatly on suffix analysis in English, and these techniques do not work, for example, in German, where suffix information can be embedded inside agglutinated compound words. Thus, a knowledge of the morphology of a language is necessary for performing multilingual spelling correction properly.

In addition to morphological considerations, there are numerous strategic variations in spelling correction algorithms that can have a significant impact on their performance. Current spelling correction systems in wide use on the ARPANET generally make an assumption that there is only one error in a misspelled word. This assumption lends itself to a convenient and inexpensive correction algorithm which in effect considers every letter in the word except one to be correct (or considers there to be only one transposition of two letters). The consequence of this simple algorithm is that often dozens of possible corrections are offered for a single-letter misspelling in a four-letter word, yet very little is done to help the user making two or more errors in spelling a 15-letter word. “Internationalization” can readily be recognized by English speakers even with several errors—“internationalisation” contains nn for n, tt for t, and s for z, yet it is readily manually corrected. Finally, spelling correction is beginning to appear in the literature of computer science as both the theoretical basis and the necessity for it are becoming better recognized.

Computational lexicology will assist spelling correction developments in three ways. By surveying full-text sources, it will be able to provide information on the frequency of spelling errors and their types. By amassing new lexicon, it will be able to provide spelling correction algorithms with improved data for their use—lexicon will contain spelling forms from special subject domains as well as increased quantities of error-free lexical and morphological information (from published dictionaries). Finally, by providing statistical information on the frequencies of morphological and phonological forms, it may be possible to design better spelling correction algorithms.

A note to the topic of spelling correction: Word hyphenation is becoming a feature of some word processing systems. Word hyphenation is poorly understood and even more poorly described in algorithmic form. Spelling correction lexicons are logically the place to incorporate hyphenation data because of the savings in storage accompanied by inserting the hyphenation points into the spelling dictionary vs. having a separate lexicon for hyphenation. Thus the field of spelling correction may soon broaden its boundaries to become the field of word correction. It would involve not only repairing misspellings, but providing correct hyphenation information and perhaps additional capabilities such as checking for abbreviation consistency, checking for capitalization, foreign language transliteration, and even font changes (e.g., italicizing foreign phrases). The work on the “Writer’s Workbench” UNIX programs at Bell Labs and the WORDS package offered by Houghton Mifflin (Houghton Mifflin, private communication, 1981) appears to be leading in the direction of developing such computational tools for correcting words in text.

Machine Translation

The current state of the art in mechanical translation (MT) is extremely hard to assess. The field is charged with highly controversial claims, international rivalries, and a historical stigma. Despite these problems, one facet of mechanical translation does become clear. It depends on a massive lexicon.

Although practitioners of MT may never resolve their arguments over computational linguistic strategy or even systems design directions (machine-assisted vs. fully-automatic) it is possible to make advances in the necessary prerequisites for mechanical translation without entering into the fray (or at least without becoming locked in unresolvable battle).

One of the conspicuous lacks in mechanical translation is the availability of an adequate lexicon. Naive MT consumers very often conceive of MT as a task involving a simple paired word list (or even worse, a multilingual word table) and an amazing and mysterious computational algorithm which can take this word list and use it to perform translations. In reality, the algorithms necessary for performing mechanical translation are within our reach today, but the grammar and lexicon these algorithms would require to provide flawless high-quality mechanical translation are many years from attainment. Linguists would readily admit that there does not exist a complete grammar of any major natural language. Lexicologists will likewise explain that a complete lexicon, with all the grammatical information needed to perform MT, is nowhere to be found.

Thus, unlike the naive view that the MT problem requires finding or building the marvelous algorithm, the actual task involves creating an adequately detailed description of a language in terms of its syntactical possibilities and lexical semantics. Computational lexicology is thus likely to be necessary in providing an MT system with its basis for operation.

Retrieval without translation

A very significant consequence of the increased availability of machine-readable full-text sources and the accompanying growing shortage of economical language translation capability is that a high premium will be placed on determining whether a document is worth translating before its translation is undertaken. Thus full-text sources in foreign languages such as Russian, German, French, and Spanish will become readily available via satellite telecommunications, which could put a foreign newspaper on our doorsteps in a matter of hours after its publication in its native country. However, the translation of this text would require a major effort—one which would not be undertaken unless potential readers knew there was something of interest to them in the text that could justify the cost of translating.

Therefore a situation is set up in which a full-text filtering system would be desired which operated directly on the foreign language full-text source material and rendered a judgment about the content before sending such material to a translator. The only real indicator of the content of a document that can be used without full grammatical parsing is an analysis based on the words used in the text. This in turn implies that a sophisticated lexicon will be the prerequisite to filtering untranslated foreign source material.
REFERENCES


