Introduction to Natural Language Processing

Mary Dee Harris

Introduction to Natural Language Processing

Mary Dee Harris
Loyola University
New Orleans

Reston Publishing Company, Inc.
A Prentice-Hall Company
Reston, Virginia

To the Reader

Natural language is any language used for communication by humans. *Introduction to Natural Language Processing* is an attempt to provide enough information about natural language and the problems involved in dealing with it to allow rudimentary natural language processing by computers and to make apparent what kinds of processing can be done now. In addition, and perhaps more significantly, what is not yet possible will be discussed. This book consolidates information about language from the literature of several disciplines: linguistics, artificial intelligence, psychology, and cognitive science, and describes the computer science tools and techniques available to deal with natural language.

From an information processing viewpoint, the language information to be processed must be conveyed from the outside world to the processor and the resulting text or utterance conveyed back to the outside world. In other words, a system to process natural language must be capable of input and output of natural language. No matter what the original form of the language: speech, printed text, or whatever, the external form must be communicated to the processor and stored in some internal form. After the processing is completed the data must be communicated back to the user—whether the form is video screen, hard copy (e.g., paper), or speech. Once the language data has been input to the system, it must be stored in a form that preserves aspects of the original and allows manipulation for whatever purpose needed. The language data must be organized into appropriate data structures which are capable of recording information at various levels: morphological, grammatical, contextual, etc. Information about the structure of the discourse and about its meaning must be stored. All of the representations of data can be referred to generally as natural language structures to correspond to the memory of the general information processing system.

Between the input of language and the output of language various manipulations will occur, depending on the purpose for which the system was designed. Processing a sentence can include various operations on a specific kind of natural language element: analyzing a sentence by breaking it into its constituent parts, translating the sentence into another language, generating a sentence to be output from the system. Any natural language unit can be manipulated. The primitive operations consistent with units at any level are: synthesis (generating the unit from smaller components); analysis (breaking the unit up into its constituent components); transformation (used in a general sense, changing the unit into an equivalent unit); and inference (drawing conclusions from the knowledge already stored).

These aspects of rendering natural language—synthesis, analysis, transformation, and inference—are necessary in order to understand natural language, to be able to use language as humans use it—for requesting information, for stating knowledge, or for restating discourse without loss of information. Systems which can do all these things can be used as question answering systems. They can generate well-formed sentences, interpret requests for information, paraphrase statements, and store knowledge provided. The amount of understanding that can be said to occur is dependent upon the size of the vocabulary used, the difficulty of the questions which are answered correctly, and the degree of subtlety and complexity of language handled.

This book has been designed for two primary audiences: undergraduate Computer Science students and programmers interested in adding natural language interfaces to their software products. As a textbook for Computer Science majors with at least two programming courses, some of the material will be review, such as the later sections of the chapter on text processing and the explanations of elementary data structures, such as stacks, linked lists, and trees, throughout the book. This material has been included for those readers who have not learned these concepts rigorously enough to understand the subject in general without at least a brief explanation. In addition the book can be read by non-technical people with little or no background in programming, if they skip the more technical sections. This book has been used twice in prepublication form for an upper-level undergraduate Computer Science course at Loyola University in New Orleans. The course covered all the material in the book, except the sections on text processing and data structures mentioned above, since the students knew those concepts. In addition to the text, the students developed four major programs based on the exercises, read additional articles from the bibliography, and wrote a term paper on some aspect of natural language processing not covered in the book. At the end of the semester the students were able to read current literature in natural language processing and cognitive science with no difficulty and were prepared to begin graduate study of a specialized area of the field.

In this book we consider the design of a natural language system (NLS) using as a paradigm the psychologists' model of humans as information processing systems. Our NLS communicates with its environment in written text. The discourse is translated into internal structures for rendering. These structures are interpreted abstractly; in other words we are concerned with the logical aspects of language and the structures inherent in it rather than machine requirements for representing a character, array, or pointer. Similarly the algorithms for rendering natural language relate to the kinds of operations to be performed on language. Defining appropriate data types and the operations upon them allows us as system designers to be concerned only with the definition and manipulation of language data, rather than the details of implementation.

The focus of this book is an attempt to understand language and the means of processing language from as many perspectives as possible, with emphasis on both analysis and synthesis techniques. The book is practically oriented in the sense that the elements of the system developed are not theoretical, and could be implemented on a computer system. However, the system is not guaranteed to deal with all problems of processing natural language. In fact, it is guaranteed not to solve all the problems because not all the problems have even been identified yet. And the techniques presented are not adequate to handle all the difficulties of understanding natural language. But natural language systems are relatively new; many people have considered the advantages of having computers capable of dealing with language, but natural language processing has not had the emphasis of many other areas of computer science. The treatment of the subject in this book is to provide some *nuts and bolts* techniques for approaching some simple problems in dealing with natural language and to explain some of the terminology and methodology used by researchers in the field. Anyone who works through this book will have a good grasp of major problems of computer *understanding* of language and should be able to apply these ideas to a restricted version of the English language. He will be able to read and understand the current literature describing research in Natural Language Processing and Cognitive Science. And hopefully, some readers will find better solutions to old problems and also uncover new problems.

The method used to describe data structures and algorithms is an algorithmic language or pseudo-code based on Pascal with several fairly standard extensions. For example, since text manipulation is facilitated by string functions, the pseudo-code will include UCSD Pascal string functions with some SNOBOL-like pattern matching functions defined in the chapter on text processing. Some of the processing operations to be discussed in Unit IV will require concurrent processes and therefore we will assume concurrent processing to be a part of our algorithmic language. The algorithms will look much like Pascal, but will in fact allow several features not in "standard" Pascal. Although much of the work in processing natural language has been done in LISP, I have chosen not to use it for this book because it is not as widely available nor as widely understood as a Pascal-like language. However, through the extensions to standard Pascal and the liberties I have taken with it, some of the pseudo-code will not always look like Pascal.

Computer science is not a spectator sport, and natural language processing is no exception. With that in mind, I have included exercises in each chapter to help the reader understand the concepts covered. To quote Donald Knuth's "Notes on the Exercises," from Volume 1 of *The Art of Computer Programming*,

> It is difficult, if not impossible, for anyone to learn a subject purely by reading about it, without applying the information to specific problems and thereby forcing himself to think about what has been read. Furthermore, we all learn best the things that we have discovered for ourselves. Therefore the exercises form a major part of this work; a definite attempt has been made to keep them as informative as possible and to select problems that are enjoyable to solve. [1]

The exercises vary in difficulty and in orientation; extending Knuth's ideas, the reader is encouraged to work as many exercises as he can stand. Many of the exercises involve writing programs (of course). All programming attempts should start with a design phase before coding begins and end with a testing phase. Not only will this procedure reinforce good programming practice, but it will help the programmer understand the limitations of the approach taken in the book. In other words, the reader can discover for himself, during the design phase, what aspects of language are being ignored, and, during the testing phase, what kinds of discourse will be interpreted incorrectly. Constantly considering the limitations may seem to be a strange approach to follow, but doing so will remind the reader that the study of natural language processing has only begun. Even the non-technical reader can benefit from attempting the exercises, by writing a design in logical and clearly stated English. Experience has shown that non-technical readers can produce designs, which look very much like pseudo-code, thus accruing all the benefits of the understanding produced by designing a piece of software.

This book is organized to emphasize three primary aspects of information processing—definition of the input and output functions, description and manipulation of the data, and design of the overall software system, including both data and program structure. Unit I presents background material on the study of language and an overview of linguistics, for the benefit of those readers who have no knowledge of the field. The chapter emphasizes an understanding of the terminology that is used in the areas of linguistics, artificial intelligence, and cognitive science.

Unit II of the book deals with input and output of natural language. Chapter 2 describes specific problems of text processing, and Chapter 3 considers the lexical phase, especially the organization and access of the dictionary to hold the vocabulary for the system. An overview of the primitive operations of manipulating text seems within the scope of this book, even though much of the coverage will be review for many readers. The lexical phase extends the discussion of text manipulation into the specific considerations of lexical analysis and generation.

Unit III, Natural Language Structures and Algorithms, covers the three general areas: linguistic structure, the definition of the relationships among linguistic elements; the correspondence between linguistic structures and the world, how to represent the meaning of language elements; and cognitive processes, involving the structure and manipulation of knowledge required for storing concepts acquired through language input. This unit can be considered the heart of the book in that it describes the structure of the natural language data, the actual operations to be performed on that data, and suggests methodologies for these operations. Moving along the spectrum from syntax to semantics to representation of knowledge and from simpler to more complex problems, this unit presents problems of synthesis, dealing specifically with generation of sentences; problems of analysis, or parsing and interpreting sentences; and problems of transformation of language elements, paraphrasing, translating, and linguistic transformation. In addition, the capability for inference and question answering is evaluated for some of the methodologies discussed.

Unit IV, Natural Language Systems, attempts to incorporate the ideas from the earlier parts of the book into a design scheme for a complete system for manipulating natural language. This system is capable of acquiring knowledge from the input, answering questions about the knowledge stored, and generating appropriate responses for output. This unit covers the problems of data storage design and the organization of processing, with emphasis on the lexicon. Traditional approaches to system design are considered as well as more workable, recent schemes. All the methods discussed are evaluated in terms of the constraints and limitations imposed. The result is an understanding of the conceptual and pragmatic problems involved in designing a system for understanding natural language.

Contents

Introduction to Natural Language Processing

Mary Dee Harris
Loyola University
New Orleans

Reston Publishing Company, Inc.
A Prentice-Hall Company
Reston, Virginia

Library of Congress Cataloging in Publication Data

Harris, Mary Dee.
Introduction to natural language processing.

Bibliography: p.
Includes index.
1. Interactive computer systems. 2. Programming
languages (Electronic computers)—Semantics. I. Title.
QA76.9.158H37 1985 001.64'2 84-18242
ISBN 0-8359-3254-0

A Prentice-Hall Company
Reston, Virginia 22090

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America

Part I

An Introduction to the Study of Language

Introduction

What is natural language? Natural language is any language that humans learn from their environment and use to communicate with each other. Whatever the form of the communication, natural languages are used to express our knowledge and emotions and to convey our responses to other people and to our surroundings. Natural languages are usually learned in early childhood from those around us. Children seem to recognize at a surprisingly early age the value of structure and uniformity in their utterances. Words, phrases, and sentences replace grunts, whines, and cries and better serve to convince others to recognize the child's needs. Natural languages can be acquired later in life through school, travel, or change in culture, but with very few exceptions, all humans in all cultures learn to communicate verbally in the language natural to their immediate environment.

In contrast to natural languages, artificial languages are languages created by humans to communicate with their technology, for example, computer programming languages. The term *artificial language* implies a language consciously crafted by humans rather than a language learned naturally and includes languages such as Esperanto, a *constructed language* and designed to be universal, easily learned by speakers of any natural language. The definition of natural language should not make excessively rigid distinctions that eliminate relevant languages. I will appeal to the reader's intuitive notions of natural language rather than attempting a rigorous definition.

Humans process natural languages whenever they read Shakespeare, dictate a business letter, or tell a joke. Sign language is used by the hearing impaired to communicate thoughts and feelings with others and replaces the language they are unable to hear. Despite the different forms of language in each of these situations, aspects of the language used are similar. Whether language is spoken or written, every message has a structure and the elements of language relate to each other in recognizable ways. Verbal communication or speech is characterized by the sounds which almost every human is capable of producing. Whether each person learns to produce a particular sound is determined by the languages learned rather than the anatomical speech production mechanisms, which are approximately the same for all normal humans. Speech is produced by stringing together individual human sounds in recognized patterns. The study of these patterns of sounds is called *phonology*. The study of the structure of language units and their relationships is called *syntax*. Phonology and syntax are both important parts of the field of linguistics.

Linguists are also concerned with semantics, the study of the relationship between the linguistic structures used and the meanings intended; in other words, how does what we say or write relate to what we mean? It is not enough for a sentence to be correct in form; it must also make sense. For example, the sentence,

The tree sang the chair.

would not in ordinary discourse, be a meaningful sentence, even though it is grammatically reasonable. The noun phrase, *The tree*, can be the subject of a sentence; *sang* is obviously a verb; and *the chair* is a noun phrase which can serve as a direct object of a verb. But trees do not sing, and nothing that sings, sings chairs. So, how can this string of words be a sentence, even an unreasonable one? Many people, including linguists, would say it cannot be considered a sentence. Imagine, for a moment, a fantasy movie set some place like Middle Earth or another planet, where trees were the intelligent beings and had what we consider human characteristics. If trees could talk in this world, then they might also be able to sing. And if we use our imagination some more, we might find that speech and song in this strange world were made of physical objects such as chairs and cups, rather than phonemes. That may seem farfetched, but sound waves are also physical and produce recognizable effects in physical objects. If that is too extreme, consider that the phrase, *the chair*, is the title of a song, and the tree is made up of Christmas carolers standing on bleachers in the shape of a Christmas tree. Does the sentence still seem to lack meaning? Could we not say under some circumstances, however strange, that the sentence, *The tree sang the chair*, makes sense?

Obviously language is complex and trying to understand language is sometimes hard for humans. People do not know very much about what it means to *understand* language, partly because they do not often think about language; they just use it. Lots of people can ride bicycles without *understanding* what makes it possible for them to be able to do so. One need not learn the physiological explanations for pedaling, steering, and balancing to get around on a bike. Physiologists know a great deal about how a person can ride a bicycle, how the muscles and ligaments interact with the eyes, the nerves, and the bicycle to produce a system that can be referred to as *riding a bicycle*. The same is true of language. Humans use language without knowing how to use language, and linguists attempt to explain the system that we call *using language*.

However, most of us know more about language than we realize. Some language researchers believe that language is an innate capability in humans, that all humans learn a language because the structure of the language is a biological aspect of the species. Studies of language acquisition seem to indicate that motor development and language development are related; most babies learn to talk at about the same time they start to walk. When babies start babbling before they use the language of their parents, their utterances frequently sound like they *should* make sense. A baby *talking* on a toy telephone, uses intonations and phrasing similar to his parents', even though the particular sounds have no recognizable meanings. In other words, the child's sounds have the structure of his parents' language without any content being associated with the sounds. Similarly we recognize that the following sentences have the same structure.

The tree sang the chair.

The students finished the exam.

The new, young vice-president in charge of financial affairs in the company established extraordinary regulations concerning the procedures for reporting exceptional situations in the payroll department.

Thus, something in language makes us aware of similarities among sentences despite the variance in subject matter. The systems used by linguists to describe these similarities are called *grammars*. The term *grammar* is also used to refer to the methods taught in school such as diagramming sentences. These methods are designed to show the relationships among the various structures within sentences. Grammars consist of the elements allowed within sentences and the rules for putting these elements together. For example, the structure of some sentences can be described as a noun phrase followed by a verb phrase. In this case, the elements are the *sentence*, a *noun phrase*, and *verb phrase*. A rule to express their relationship could be written as:

SENTENCE → NOUN PHRASE + VERB PHRASE

Obviously further definition of these elements would be required to describe noun phrases, verb phrases, and their components, and each of the components would have to be defined. This process of redefinition of the grammatical constructs would be continued until the elements were defined as specific words. The words in a grammar are called the *vocabulary*. The grammar is made up of the rules and the vocabulary along with the meanings associated with the vocabulary. Various grammars will be considered in detail throughout this book.

Besides linguists' use of grammars for describing language, grammars are used by logicians and formal philosophers to study formal languages. A formal grammar is essentially a set of rules and a list of elements upon which the rules can be applied. In natural language; the elements would be words and phrases, and the rules would specify how the elements can be combined. Formal languages are composed of sentences generated by a formal grammar, and provide a means to study the features of the language. Attempts to create formal grammars of natural language have produced many, but not all, of the forms actually used in that natural language. The logician Richard Montague did not make a strong distinction between formal and natural languages, but rather claimed that a subset of English could be considered a formal language. He wrote that "the syntax, semantics, and pragmatics of natural language are branches of mathematics, not of psychology," and the syntax of English is as much a part of mathematics as number theory or geometry. This view is a corollary of Montague's strategy of studying natural languages by means of the same techniques used in metamathematics to study formal languages. "... Generalizing [metamathematics] to comprehend natural language does not render it any less a mathematical discipline." [1] Montague created several grammars to define fragments of English. One of his grammars first defined the lexicon which specifies the basic expressions in the language, words, phrases, etc. It then gave the rules to combine the expressions from the various categories to build new expressions. Montague considered the meaning in the language to be related to the structural aspects. To quote Montague, "MEANINGS are functions of two arguments—a possible world and a context of use.... Meanings are those entities that serve as interpretations of expressions...."

Other logicians have sought to represent natural language by means of propositional logic. All sentences in the language considered are written as propositions and can be manipulated according to the rules of formal logic. The statement, *All teenagers drive cars,* could be rewritten in logical notation as:

FOR-ALL (x)(EXISTS(y)(TEENAGER(x) → (CAR(y) AND DRIVE(x,y))))

When a sentence has been thus transcribed, the rules of logic can be applied to test the validity of any references to the information contained in the sentence. However, propositional logic can express only a subset of all sentences in any natural language. The method only applies to statements about which the truth or falsity can be known. Other types of logic, such as modal logic and clausal logic, can handle sentences such as the examples below dealing with fantasy or conjecture:

Hobbits can be recognized by the dark fur on the tops of their feet.

John believed he would discover the fountain of youth despite many years of failure.

John believed he would be promoted despite his boss' refusal to discuss the matter.

Logicians can handle many instances of natural language, but their methods often must rely on restricting the language to certain types of statements. Researchers are continuing to derive methods for representing various types of sentences.

Cognitive psychologists, concerned with how humans think, approach language from a different perspective than linguists and logicians. They view language as a representational medium for thought rather than viewing language as an independent phenomenon.

> Any form of learning or memory involves some form of internal representation of past events. A set of internal representations is undoubtedly a set of symbols. The calculus of operations used to relate and manipulate symbols both to each other and to the external world is a system of thought. Language is a shared community of externalised symbols, and intelligence a comparative term that we use to describe our attempts to relate the thinking power of one organism to that of another. [2]

This view of language is derived largely from the notion of humans as general information processing systems. Humans can receive signals through their senses, can discriminate between trivial and important information, can remember and categorize many pieces of information collected from many varied sources, and communicate this information back to the external world through speech or other means. Some terms used to describe the [illegible] of such systems include *receptors* to obtain signals from the world external to the system, a *processor* to manipulate the information received, a *memory* to store information, and *effectors* to convey signals to the external world. Some of the terminology reinforces the analogy between human general information processing systems and *digital general information processing systems*, more commonly called computers. Some researchers argue that human information processing systems are devices with only a single channel for input and output, with a processor of limited capacity, while others believe that the human system is a *multichannel multiprocessor device* like computer systems developed in more recent years. [3] Whichever opinion one accepts, the notion of human intelligence being explained as an information processing system provides new insights into the field of cognitive psychology and will serve as a useful model for our discussion of processing natural language.

Other psychologists consider language not so much in terms of what it reveals about our thought processes, but rather in light of how it represents social interaction. We learn language from the people we are close to as children, and we talk to each other for emotional support and cooperation as well as for intellectual communication. A large percentage of all verbal utterances communicate emotional and social needs more than intellectual content. Written language is less often socially oriented than spoken language because the purpose of writing is generally to record ideas, rather than feelings;

however, writing personal letters is a good example of social communication. Language used for social purposes follows the same rules as other language, but frequently is highly formulaic. The same phrases are used over and over in similar situations, often losing their meaning somewhat, yet still serving their basic function. For example,

> I love you.
>
> Hello, how are you? Fine, thank you, and you?
>
> Thank you very much for the gift. I like it a lot.
>
> Hey, bro. Wha's hap'nin'?

Dealing with language of this sort requires different analysis techniques from other language. The meaning behind the words seems to be of a different nature than the content of language used to convey information. Yet the notion of language as social interaction still fits the paradigm of the human information processing system.

TX/eft 733

Another approach to defining and studying natural language is in the area known as Cognitive Science, a relatively new field combining knowledge and research methods from a number of other fields: computer science, particularly artificial intelligence; mathematics; psychology; and linguistics. In the first issue of the journal *Cognitive Science*, one of the editors, Allan Collins, attempts to answer the question, "Why Cognitive Science?"

> Cognitive science is defined principally by the set of problems it addresses and the set of tools it uses. The most immediate problem areas are representation of knowledge, language understanding, image understanding, question answering, inference, learning, problem solving, and planning.... The tools of cognitive science consist of a set of analysis techniques and a set of theoretical formalisms.... Unlike psychology or linguistics which are analytical sciences and AI which is a synthetic science, cognitive science strives for a balance between analysis and synthesis. [4]

Cognitive scientists do not view language as their primary area of concern, but, like the cognitive psychologists, approach language as a representation of thought and memory which is shared among some group of people. This sharing allows for communication by means of the language. Thus, language is not a separate and distinct concept or entity, and it cannot be understood without information about the context of the specific instance of language use and the background of the person using the language. Terry Winograd, well-known for his work in artificial intelligence, presents four *phenomenic domains* for understanding language in the article, "What Does It Mean to Understand Language?"

> *The domain of linguistic structure* which is concerned with the structural elements of language.

> *The domain of correspondence between linguistic structures and the world*; in other words, what do the structural elements of language refer to, or "mean".
>
> *The domain of cognitive processes* which involves the structure of knowledge and the manipulation of the items in the structure by the processor of the language (either human or computer).
>
> *The domain of human action and interaction* which views language within time, relative to past language use and future expectations.[5]

These four domains recall the various views of language by several of the disciplines considered so far. The domain of linguistic structure has traditionally been the purview of linguistics, and has more recently been approached by the logicians. Psychology, linguistics, and philosophy have studied the domain of correspondence between linguistic structure and the world. Cognitive psychologists and biologists have investigated the cognitive processes and mechanisms involved in using language. And psychologists have always been concerned with the domain of human action and interaction involving any form of communication. Winograd believes that the field of cognitive science should attempt to integrate all these approaches and to use any of the domains required to extend their capabilities of understanding language.

In another article in the first issue of *Cognitive Science*, "Artificial Intelligence, Language, and the Study of Knowledge," Ira Goldstein and Seymour Papert discuss "the relationship of Artificial Intelligence to the study of language and the representation of the underlying knowledge which supports the comprehension process." They argue that AI has shifted its perspective from a "power-based strategy for achieving intelligence to a knowledge-based approach." [6] Quoting Minsky and Papert, they point out that,

> The *power* strategy seeks a generalized increase in computational power. It may look toward new kinds of computers . . . or it may look toward extensions of deductive generality, or information retrieval, or search algorithms, . . . In each case the improvement sought is intended to be "uniform"—independent of the particular data base.
>
> The *knowledge* strategy sees progress as coming from better ways to express, recognize, and use diverse and particular forms of knowledge. This theory sees the problem as epistemological rather than as a matter of computational power or mathematical generality. It supposes, for example, that when a scientist solves a new problem, he engages a highly organized structure of especially appropriate facts, models, analogies, planning mechanisms, self-discipline procedures, etc. To be sure, he also engages "general" problem-solving schemata but it is by no means obvious that very smart people are that way directly because of the superior power of their general methods. [7]