Preface

This is a book about building computer programs that parse (analyze, "diagram") sentences of English the way you learned to (or wish you had or are thrilled you didn't) in school. We are talking about "real-world" English, quotidian prose—say the entire text of any book in your local library concerning needlework; or of any wholesale footwear catalog; or of any physics journal; or of this morning's newspaper.

Looked at abstractly, the problem of parsing quotidian English by machine is one of "event prediction". The evidence you have to go on is the sequence of words forming the sentence for analysis. Your first step is to identify every possible parse, every *possible* construal of the sentence, however unlikely it is that any sensible person could have meant *that*. Second, and finally, you need to choose among all the merely possible breakdowns of the sentence the one that was *most probably* intended. So, given a pool of "events"—all the parses you recognize as potentially valid for a given sentence—your parsing program engages in "event prediction" by choosing the "right" (the intended) analysis, just as you did (or didn't do, or...) in school.

The programs that do exist for this purpose, and there are many of them around the world, are actually not very good even at establishing the event pool for everyday sentences, let alone at choosing the intended event (the "intended parse"). For example, any page of this morning's newspaper contains sentences for which most parsing systems today cannot offer any analysis at all—the systems would be stumped. And that same page contains a lot more sentences that most parsing systems would not be able to analyze *correctly*, even where the definition of "correctness" is very forgiving indeed. In Chapter 1 we adduce evidence to back up the above claims.

It turns out that precisely those research domains where the description of the "event space" is not sharp, not comprehensive, not reliable—that these are the research areas that have proven to be the most fertile ground for the application of statistical methods. The outstanding example of this principle in recent years has been the field of "acoustic modelling" for speech recognition. Here the task is to identify the speech sound intended by a speaker, on the basis of the sounds he or she has uttered in the course of speaking. The scientific vocabulary available for the description of the speech sounds themselves is not dependable, and it has worked out over the last five years or so that the field as a whole has converted to a statistical approach to the acoustic modelling problem—that is, the vast majority of published papers in this area now treat of statistically-based research. And the result has been a very marked improvement in the accuracy of these

identifications. As a matter of fact, the particular statistical algorithms which have spelled success for the acoustic modelling field are the same ones that we put forward and explain in this book. Other research domains which are currently in the process of switching over to a statistical basis are image processing and underwater acoustics.[1]

Three approaches, broadly speaking, have been tried to date to the problem of building computer programs that parse English sentences. The differences among these approaches have to do with the two tasks involved in producing a computer program that correctly parses English, and whether each task is attempted by a human expert or via statistical data processing. Recall that the two tasks are the task of identifying every possible parse, and the task of choosing the right, the intended parse among all the merely possible parses.

Perhaps surprisingly, the method of attack which has been most favored among researchers is the "artisanal" approach: the human—i.e. the grammarian—specifies rules or procedures for both tasks. That is, he or she "dreams up" not only the rules for diagramming English sentences, but also the procedures for deciding when one possible parse is more appropriate than another.

A radically different means of broaching the parsing problem is the fully automatic method: Statistical methods are directed toward the solution of both the subtasks. On this approach, a "grammar" or set of rules for analyzing English grammatically must be "discovered" in some fashion by statistical algorithm, and then these rules need to be applied in the optimal way, again using statistics, so that the correct parse is always found from among the set of available parses. Work in the direction of this fully-automatic analysis strategy has begun, but is still in the exploratory stage. Pioneering research was done by Sharman ([SJM88]), and more recent work has included [PS92], among others. Moreover, a fairly intensive effort is underway at this time at the IBM T. J. Watson Research Center in Hawthorne, New York, but this research has not yet been reported on in print.

Finally, and most relevantly to our purposes here, there is the approach of allowing the human expert, the grammarian, to come up with the rules for "diagramming" English, but then loosing the power of statistical methods on the more complex task of deciding what factors influence the choice among multiple parses for a given sentence, and precisely how the factors

[1] Image processing is the reconstitution of visual images and the enhancement of imperfect ("noisy") visual images such that they become better defined ("cleaner"). Underwater acoustics involves inferring characteristics of underwater objects through the analysis of sound waves with which they have come into contact.

interact in conducing to a final decision as to which parse is "best".[2]

This book is about this "mixed" approach to the problem of parsing English. The grammarian *supplies* the rules of linguistic analysis and the statistical algorithm *applies* them to English sentences of the sort we find all around us every day.

Our book has two aims. It is a "how-to" book, on the one hand, and as such it aims to show you how to build a statistically-driven broad-coverage grammar of English. We even supply you with our own grammar, specified in detail in Chapter 4; with the necessary statistical algorithms, which are set forth in Chapter 6 and, in a different way, in Chapter 7; and with the requisite knowledge, in Chapters 2 and 3, for you to prepare whatever set of English-language sentences you wish to work with, so that they can be used to guide the statistical process in applying the grammar's rules.

But at the same time our book is a record of a five-year-long collaboration between the Continuous Speech Recognition Group of the IBM T. J. Watson Research Center, Hawthorne, New York, USA, and the Unit for Computer Research on the English Language (UCREL), University of Lancaster, UK. UCREL has been in the business of preparing very large amounts of data—of everyday English sentences—to be parsed by a human-created program for grammatical analysis (a *grammar*) and processed by statistical methods so as to direct the grammar, so to speak, towards the most likely parse, among all those offered up, for any input sentence of English. The IBM Continuous Speech Recognition Group's team has been in the business of creating the needed grammar and refining and applying the appropriate statistical routines.

While we at UCREL and at IBM Research have been working together now for five years, we will be reporting most closely on our efforts of the past two years or so to apply our statistical approaches to the task of automatically obtaining the correct parse for any sentence of any *computer manual.* The three years prior to this most recent effort saw much productive research on our part, we believe, but also, and perhaps more importantly for the purposes of the present book, saw us make many mistakes. For instance, we aimed too high at first in attempting to parse a very "general" variety of English, that of the Associated Press news wire. Or again, we made what was perhaps the converse mistake (see Chapters 2 and 5) of preparing our data in Lancaster in exact conformance with (a precursor of) IBM's grammar, so that normal changes in our grammar tended to diminish the usefulness of the processed data by rendering it partially obsolete.

[2]The fourth possible combination of human and statistical efforts—the one where the machine composes the rules for analysis (the grammar), and the human dictates how the rules are to be applied to sentences for analysis—this approach has never been tried, to the best of our knowledge.

The point is that we have spent the time and money involved in making and working through these mistakes. Our goal, our assumption, and our hope is that you will therefore not need to suffer through the same lapses in progress that we did. Of course, you will probably encounter other pitfalls, but at least these will be new to the field, and after all, research—and product development, and, of course, any endeavour—involves a constant learning both from one's errors and from one's successes.

So the *relevance* of our research experience, as reported to you here, is not merely academic. It dovetails with the hands-on focus of the book in that it offers you a case history of the application of the principles and methods we try to impart to you throughout.

We have made a point of trying to address this book to the broadest possible readership. We have had in mind really anyone with an interest in the field of machine grammatical analysis, but most particularly linguists, practical and theoretical, computer scientists, workers in artificial intelligence, programmers, mathematicians, and last but not least, students or managers in the above fields. We feel this book lends itself to use in university courses or in industry, where there is an interest in or a need for accurate automatic parsing of English. The mathematics of Chapter 6, and to some extent Chapter 7, does require some mathematical sophistication, and here we have imagined the reader to have an undergraduate- or graduate-level computer science background. But it is very possible to understand our methods and approach, in a general sense, without fully grasping the mathematics of these portions of the book. Moreover, we have worked as a team—some of us are mathematicians, some of us are computer scientists, and some are grammarians. An approach to this book involving two or more different readers cooperating on the task of creating a statistically-based grammar of English is another useful model.

This book itself is a product of teamwork. You may notice minor shifts of perspective, and of writing style, as you read through our book. We believe that the team approach has helped our research by allowing each of us to concentrate fully on a single aspect of the difficult problem of parsing English by computer.

Before we go on to Chapter 1, and to the rest of the book, we wish to thank the following colleagues and friends who have read all or part of our book while it was still a-bornin', and who have provided welcome and useful commentary to the authors: Adam Berger, Barbara Gates, Phil Harrison, Jennifer Lai, Krishna Nathan, Adwait Ratnaparkhi, Paul Rayson, Jeff Reynar, Stephanie Smolinsky. Still other colleagues and friends—Tom Barey, Louise Denmark, Jean Forrest, Janice Hurst, Xunfeng Xu—contributed in Lancaster to some of the research we are presenting here, and for this we are grateful to them. Finally, our thanks go to Richard Sharman of IBM

UK, for his many contributions to the collaborative research we are about to describe.