

Introduction

Language study and the computer

This book charts the emergence of a new view of language, and the technology associated with it. Over the last ten years, computers have been through several generations, and the analysis of language has developed out of all recognition.

The big difference has been the availability of data. The tradition of linguistics has been limited to what a single individual could experience and remember. Instrumentation was confined to the boffin end of phonetics research, and there was virtually no indirect observation or measurement. The situation was similar to that of the physical sciences some 250 years ago.

Starved of adequate data, linguistics languished—indeed it became almost totally introverted. It became fashionable to look inwards to the mind rather than outwards to society. Intuition was the key, and the similarity of language structure to various formal models was emphasized. The communicative role of language was hardly referred to.

Although linguistics is gradually coming into balance again, it has left problems behind. Its most inward-looking period coincided with a surge of interest in computing, and a very limited type of computational linguistics became fashionable and remains the orthodoxy in many places.

This book offers an alternative. Throughout the decade of research reported, the basic method remains unchanged. Large quantities of ‘raw’ text are processed directly in order to present the researcher with objective evidence.

Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular.

The Cobuild project

Ten years ago, the English Language Research group at the University of Birmingham teamed up with Collins publishers, to investigate this

Introduction

area and produce language reference works. During the 1970s, computational research on English had not progressed much in Birmingham because all the energy was spent on preparing for the future—devising software packages, instituting undergraduate courses, and influencing opinions on the campus. In particular, there was need to stress the growing importance of data-processing at a time when computing was almost confined to number crunching.

From 1980 to 1986, this essential preparatory work was put to good effect, and transformed through experience into a completely new set of techniques for language observation, analysis, and recording. The details are charted in a book called *Looking Up* (Sinclair, ed., 1988). A huge database of annotated examples was assembled, and a substantial dictionary edited from that (Sinclair *et al.*, 1987). Arising from another strand of research in Birmingham, the study of interactive discourse, innovations were made in the defining style, but the major novelty was the recording of completely new evidence about how the language is used.

The initial aims were modest, and no one anticipated that the project would have such a wide-ranging effect. The computer was thought of as having principally a clerical role in lexicography—reducing the labour of sorting and filing and examining very large amounts of English in a short period of time. In addition, the management of a long and detailed dictionary text made demands on conventional methods of book production, and in the late 1970s the prospects of computerized typesetting were growing more realistic.

It was not expected that the project would turn up anything very controversial. English, as the world's most described language, would not be likely to provide startling new evidence that had been overlooked for centuries; no provision was made in the plans for launching a new perspective on description.

Nevertheless, the design of the database was cautious, and the categories used to collect information were as neutral as could be devised at the time. Compilers were trained to note down the preponderant features that they observed, by selecting one or more typical examples and drawing attention to the critical features in each example.

As the evidence started to accumulate, it became clear that the accepted categories of lexicography were not suitable; the foundations of a new approach evolved during the dictionary project and have been supplemented since; these are outlined in the position statement below.

Introduction

Three major areas of language patterning, besides grammar, could not be comprehensively presented in a dictionary, even an innovative one like Cobuild. These are *collocation*, *semantics*, and *pragmatics*. Collocation is the subject of current research, continuing a personal enthusiasm of the author. Semantic relationships are sketched out in the Cobuild dictionary, but the field is open for research that is not burdened with too many preconceptions. Pragmatics is often impossible to describe in relation to individual words, and needs separate research and treatment.

Cobuild has opened up a large range of research lines in the study and teaching of languages. The numerical and statistical side has hardly begun (though see Phillips 1989). Applications to language teaching through a lexical syllabus (Sinclair and Renouf 1988; Willis 1990) are now available and there is endless variety of detailed investigation to be done in the independent but parallel tradition established by Mindt (1988). Bilingual and multilingual research becomes possible as comparable corpora in other languages become available, notably in Italian, German, and Swedish, and the support of the Council of Europe in pioneering work is gratefully acknowledged. Experiments in new types of bilingual dictionary are in progress involving Brazilian Portuguese, and Arabic.

At the heart of all this activity lie a number of questions whose answers require reflection. The picture of language coming through is in crucial ways unlike what was expected. Is it wise to divide language patterning into grammar and something else (be it lexis or semantics or both) before considering the possibility of co-ordinated choice? Should we have confidence even in the assumption that discrete units of text, such as words, can be reliably associated with units of meaning? How do we represent the massive redundancy of language, which is often asserted but does not appear prominently in popular models of language?

These are fairly fundamental questions, which suggest that we should not jump into new theoretical positions until a great deal more research has been done, using the powerful new tools at our disposal.

Position statement

It is clear from the above that any position statement at the present time must be regarded as provisional. Nevertheless, it is worth making a statement every few years in times of swift change, so that the movement of ideas can be charted.

Introduction

Accept the evidence

First and foremost, the ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before. The regularities of pattern are sometimes spectacular and, to balance the variation seems endless. The raw frequency of differing language events has a powerful influence on evaluation.

The comprehensive nature of simple retrieval systems is an excellent feature. No instance is overlooked, and the main features of usage are generally clear. Minor patterns remain in the background. Some very common usages are often not featured in descriptions because they are so humdrum and routine; this method brings them to the fore. Especially in lexicography, there is a marked contrast between the data collected by computer and that collected by human readers exercising their judgement on what should or should not be selected for inclusion in a dictionary.

Indeed, the contrast exposed between the impressions of language detail noted by people, and the evidence compiled objectively from texts is huge and systematic. It leads one to suppose that human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use the language. Students of linguistics over many years have been urged to rely heavily on their intuitions and to prefer their intuitions to actual text where there was some discrepancy. Their study has, therefore, been more about intuition than about language. It is not the purpose of this work to denigrate intuition—far from it. The way a person conceptualizes language and expresses this conceptualization is of great importance and interest precisely *because* it is not in accordance with the newly observed facts of usage.

Reflect the evidence

The integrity of actual instances is a matter that has grown in importance during the period reported in this book. Most descriptive linguists with any field experience are disposed to record examples carefully and are cautious about accepting words, phrases, or sentences which have not been attested. However plausible an invented example might be, it cannot be offered as a genuine instance of language in use.

In the climate of the 1970s, this kind of position seemed pedantic, but from the very beginning of the research it was established that

Introduction

examples were not to be tampered with, and that every observation about the language was to be accompanied by at least one clear example of it. Gradually a mood of considerable humility developed, as it was realized how intricately constructed each example was. Even what seemed to be innocuous changes prompted by the need to clarify a point, led to the resultant adaptation being challenged.

There are, of course, plenty of bizarre and unrepresentative instances in any corpus—in fact, one really enlivening feature of corpus study is the individuality of examples. There are also instances which do not easily detach from their contexts, or which require a very extensive stretch of text to avoid distortion. The problem of finding and identifying typical examples is a particular research focus at present. However, the difficulties should not be allowed to support the absurd notion that invented examples can actually represent the language better than real ones.

This stance with respect to real examples still appears to be controversial. For many applied linguists, to abandon the practice of inventing or adapting examples would mean a big change; the demise of cherished methods and the wholesale revision of many publications. There is evidence now that re-assessment is beginning. In time, it will be realized that there is just no reason or motivation to invent an example when one is knee-deep in actual instances (and in the days of paper print-out the phrase ‘knee-deep’ was not always figurative!)

Natural language

What is more, the growing respect for real examples led in the mid-1980s to a notion of textual well-formedness, which was dubbed *naturalness* (Sinclair 1984). Any instance of language depends on its surrounding context. The details of choice shown in any segment of a text depend—some of them—on choices made elsewhere in the text, and so no example is ever complete unless it is a whole text. Invented examples would, therefore, appeal for their authenticity to a non-existent context, which would eventually be evaluated by someone’s intuition, with all the misleading consequences of that.

The position of those who like to invent examples would be more plausible if, in practice, it was felt that they did a good job of simulation. However, it seems that sensitivity to context is very difficult to achieve, and even experts at simulating natural language are

Introduction

prone to offer examples which are extremely unlikely ever to occur in speech or writing.

If we accept that the requirements of coherence and communicative effectiveness shape a text in many subtle ways, the term *naturalness* is simply a cover term for the constraints that determine the precise relationship of any fragment of text with the surrounding text. It is again a position of some humility. Until a great deal more research has been done, we will not know exactly what naturalness relations consist of. Until then, we should be very careful not to misrepresent a language, and in particular, we should never offer as an instance of language in use, some combination of words which we cannot attest in usage.

The term naturalness is chosen to be deliberately contentious, because the phrase 'natural language processing' has for some years been used to describe a branch of computer science which did not feature the language of naturally-occurring texts. Currently, there are signs of a growing recognition that the comprehensive study of language must be based on textual evidence. One does not study all of botany by making artificial flowers.

Units of meaning

Our appreciation of the relation of meaning to form has developed considerably in the last decade. It first of all seemed rather a coincidence that very often a particular grammatical or lexical choice was found to correlate with one meaning of a word rather than another. Perhaps a verb tense, a word order, a prepositional choice, or a collocation would be there in a large number of cases.

Students of grammar are often victims of the 'all or nothing' argument, which does not allow a few exceptions to a pronounced tendency. Students of lexis in the early days were made to feel that this kind of statistical evidence was somehow not as good as the wholesome, contrived rules of grammar. Now it is manifest that the nature of text is not to follow clear-cut rules, but to enjoy great flexibility and innovation.

This is leading to wholesale changes in the idiom of language description. In the relation of form and meaning, it became clear that in all cases so far examined, each meaning can be associated with a distinctive formal patterning. So regular is this that in due course we may see formal patterns being used overtly as criteria for analysing meaning, which is a more secure and less eccentric position for a

Introduction

discipline which aspires to scientific seriousness. Once again, the stranglehold of intuition is being relaxed.

The models of meaning that we are 'given' by linguistic tradition are the dictionary and the thesaurus. The traditional dictionary cheerfully represents words as often having several discrete meanings, but gives no help whatever as to how in practice the language user distinguishes among them—how a writer can be fairly sure that the meaning he wants to signal is the one which will be understood, and vice versa.

The thesaurus operates in a different way, grouping together words which share similar meanings. Its organization is entirely abstract and conceptual (and mysterious to most users). No justification is given for groupings, and hardly any discrimination is made among members of long lists of words and phrases which ostensibly mean the same thing.

The recognition that form is often in alignment with meaning was an important step, and one that cut across the received orthodoxy of the explanation of meaning. Soon it was realized that form could actually be a determiner of meaning, and a causal connection was postulated, inviting arguments from form to meaning. Then a conceptual adjustment was made, with the realization that the choice of a meaning, anywhere in a text, must have a profound effect on the surrounding choices. It would be futile to imagine otherwise. There is ultimately no distinction between form and meaning.

From this kind of progression, there is no going back. The dictionary must adapt to new criteria of explicitness and accountability to evidence, and the thesaurus is due for complete overhaul and redesign.

Working with lexicographers made it clear that there are no objective criteria available for the analysis of meaning, and that in practice the observable facts of usage tend to be undervalued. Specific training in perceiving the patterns of language is required, and the example of the computer is valuable in stressing the difference between the text as a physical object, and the way it is perceived by a language user.

More recently, the whole idea of discrete units of meaning is called into question by new evidence. No doubt, a new kind of discrete or at least discernible unit will emerge from this re-examination, possibly more abstract than the kind of unit that linguists are accustomed to. At present, the mood is somewhat negative, because once again some long-held points of view are coming under attack.

One such point of view is in the area of morphology, and the process

Introduction

of lemmatization. It is now possible to compare the usage patterns of, for example, all the forms of a verb, and from this to conclude that they are often very different one from another. There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity.

The other is in the vexed area of idiom and lexical items which apparently consist of more than one word. No reliable method has yet been found for circumscribing these and isolating them from their context, which is the first prerequisite for conventional linguistic description. The evidence set out in this book continually stresses the intricate patterns that knit a text together. The role of grammatical choices as indicating meaning is becoming more and more suspect.

A new perspective, and one which fits in with the direction of current speculation, is the following: decisions about meaning are made initially at a very abstract level, and also in very broad and general terms. At that point there is no distinction between meaning and strategy. A new-born communicative intent passes through various stages of realization, during which decisions about expression begin to be taken. These have lexical and grammatical ramifications, and are moved towards final form through a series of default options, unless a specific effect is specified in the design. The default options embody the rules of grammar (and the less explicit but very obvious rules of collocation). Grammar is thus part of the management of text rather than the focus of the meaning-creation.

In this view of language, emphasis is placed on large and fairly inaccessible decisions about topics, message design, and strategies, and it is hoped that it will stimulate a new wave of theoretical studies. It has never been anticipated that a close study of text will solve the problems of description, but merely that it will indicate more clearly what problems there are to solve. The challenge to speculation and abstract inventiveness is growing as our ability to organize the data becomes more secure.

All in all, it looks as if linguistics has concentrated on much too narrow an area of study. First of all, by leapfrogging many questions concerning the physical manifestations of language (leaving phonetics and speech technology some serious problems) linguistics becomes too abstract too quickly. Second, by trying to account for all the meaning in language at levels only one step more abstract than the initial step

Introduction

that establishes the units, linguistics remains not abstract enough. Third, by working upwards from very small units like phonemes and letters, linguistics hardly ever gets to whole texts of any length and complexity, and where it does it seems unable to maintain connection between the large units and the small ones.

Guide to the contents of this book

The first two chapters give guidance and advice on the practicalities of getting started in corpus linguistics. The foundation of this technique is the corpus, and Chapter 1 deals in general with corpus creation. Particular efforts in building corpora are referred to in the Bibliography. The results are only as good as the corpus, and we are at a very primitive stage of understanding the character of corpora and the relation between decisions on the constitution of the corpus and information about the language derived from the corpus.

In the early 1980s, as the multi-million word corpus became available for study, it became clear that the whole idea of a corpus of finite size was flawed. Any corpus is such a tiny sample of a language in use that there can be little finality in the statistics. Even the projected billion-word corpora of the 1990s will show remarkably sparse information about most of a very large word list.

The idea of a *monitor corpus* was born. Sources of language text in electronic form would be fed on a daily basis across filters which retrieve evidence as necessary. The proposal for a monitor corpus is expounded in Chapter 1; the first monitor corpus is taking shape in Birmingham as I write.

Chapter 2 goes into the basic processing in some detail. There are now some quite powerful concordancing packages becoming available for domestic computers, and it is valuable for users to have an understanding of how they work and what the main options are. The chapter articulates a demonstration package put together by one of the pioneers of text processing in the East—Professor Yang Hui-Zhong of Shanghai.

The central chapters of the book together report a series of studies of concordances, which in various ways, show how the corpus evidence can stimulate new linguistic hypotheses.

An early speculation was that the correspondence between observable patterns and distinctions of meaning and use might be much

Introduction

greater than was generally supposed. The likelihood of this was initially ridiculed, but gradually evidence accumulated. In Chapter 3, the proposal is put in its simple form; in Chapter 4 an analysis is attempted of a word that offers many subtly different meanings, and Chapter 5 combines a general discussion of phrasal verbs with a study of *set* which moves towards an explanation of the interaction between word-choice and context.

The movement from vocabulary words such as *decline* and *yield* to common, hardly lexical words like *set* and *in*, continues in Chapter 6. This includes a consideration of how grammars will be affected by the new evidence. Early, rather simplistic speculations about large corpora took the line that since grammar dealt with frequently repeated events involving very common words, it would be well served by fairly modest corpora. The big numbers were only necessary for capturing rare lexis.

There are all sorts of reasons why this position is untenable, and in Chapter 6 a new method for moving from particular to general is proposed and illustrated. The Cobuild grammar offers a broad picture of English grammar (Sinclair, Fox *et al.* 1990). These studies raise serious matters about the treatment of individuality and generality in this kind of language study. Every instance is unique, and yet contributes something towards the total picture.

Chapter 7 points out that most of the textual evidence is actually problematic, and raises the intriguing question of evaluation—some instances are better for some purposes than others.

The distillation of the typical behaviour of a word—its collocations—is at the centre of this research. The Bibliography will trace the development of the notion that a language has a viable and interesting lexical structure based upon the tendencies of words to attract and repel each other—for whatever reason. This is my own research thrust, and after many years of patient inching forward, it is a real pleasure to report progress in Chapter 8. Work is now advanced on a first dictionary of collocations (Sinclair *et al.* forthcoming).

The final chapter is (as all final chapters should be) a beginning rather than an end. Two of the main ideas that have arisen from this research are:

- a. the way we use language about language is much more important than is usually allowed for;

Introduction

- b. if a dictionary definition is written in ordinary English, all the subtle inferences and implications can be harnessed to improve the definition.

Expertise in the structure of language is now becoming central in information science, and grammars and dictionaries are as important for machines as for human beings. I have begun to try to understand the structure of ordinary language about language, and Chapter 9 is the first report on this.

The book is only a very small selection of the material produced in this exciting decade. But it makes a coherent account of the developments, and brings us right up to the present day.