Chapter 2

THOUGHTS ON TWO DECADES OF DRAWING TREES

Geoffrey Sampson
School of Cognitive & Computing Sciences,
University of Sussex, Falmer, BRIGHTON BN1 9QH, ENGLAND
geoffs@cogs.susx.ac.uk

Abstract The task of producing consistent, comprehensive structural annotations for real-life written and spoken usage teaches lessons that run counter to some of the assumptions of recent linguistics. It is not easy to believe that a natural language is a well-defined system, or that questions about the analysis of particular turs of phrase always have "right answers". Computational linguistics has been at risk of repeating mistakes made by the general fiels of computing in the 1960s; we need to learn from the discipline of software engineering. On the other hand, annotated corpora of real-life usage are already yielding findings about human nature that were unsuspected before these resources become available.

Keywords: English, Treebank, Susanne Corpus, Christine Corpus

1. HISTORICAL BACKGROUND

If one lives in the English countryside, now and then an aeroplane flies over and photographs one's house, and then someone calls to sell the picture as a souvenir. On the wall of my study I have a picture taken from the air in the summer of 1983; if you look very closely you can see in the garden a white disc, with a much smaller pink disc adjacent to it. The pink disc is the top of my bald head, and the white disc is a garden table covered with papers, because the photographer happened to record one of the opening days of my career as a producer of natural-language parse trees. I was working then at Lancaster University, where my senior colleague, Geoffrey Leech, had a new research project concerned with statistics-based automatic parsing, and I had undertaken to parse manually a sample of written English to serve as a source of statistical data for the project. Glancing at the picture since, I have often

wondered how happy I would have felt about embarking on that task, if I had known how many years of my life I was destined to devote to it.

That original "treebank" - structurally analysed sample of natural language[1], – was for internal project use and was not published; but it led me on to develop the SUSANNE Corpus, another treebank of written English, which has been publicly available since 1992 and has established itself as a popular language-engineering research resource, and more recently the CHRISTINE treebank of spoken English, the first stage of which was circulated in summer 1999. Shortly before the Paris treebank conference, I was given the go-ahead for a new project, "LUCY", to create a treebank of written English with special relevance for the teaching of writing skills. When LUCY is completed, I shall have been working on English structural annotation, with the help of various researchers at different times, more or less continuously for the two decades of my title[2].

Before looking at details of this work, it is worth reminding ourselves of the flavour of computational linguistics as it was when I began drawing trees. In those days, most computational linguists did not work with real-life corpus data, and did not seem to want to. They made their data up out of their heads. By coincidence, 1983, the year when I began drawing trees, was also the year of the inaugural meeting of the recently-founded European Chapter of the Association for Computational Linguistics, and that conference (held at Pisa) was a good opportunity to take stock of the current state of play in the subject[3]. Here is a typical selection of example-sentences used by speakers at the Pisa meeting to illustrate the workings of their systems:

Whatever is linguistic is interesting.

A ticket was bought by every man.

The man with the telescope and the umbrella kicked the ball.

Hans bekommt von dieser Frau ein Buch.

John and Bill went to Pisa. They delivered a paper.

Maria é andata a Roma con Anna.

Are you going to travel this summer? Yes, to Sicily.

By contrast, here is a sample of utterances from the CHRISTINE treebank, based on material representing spontaneous British speech of the 1990s collected in the British National Corpus:

well you want to nip over there and see what they come on on the roll

can we put erm New Kids # no not New Kids Wall Of # you know

> *well it was Gillian and # and # erm* {pause} *and Ronald's sister erm* {pause} *and then er* {pause} *a week ago last night erm* {pause} *Jean and I went to the Lyceum together to see* Arsenic and Old Lace
>
> *lathered up, started to shave* {unclear} {pause} *when I come to clean it there weren't a bloody blade in, the bastards had pinched it*
>
> *but er* {pause} *I don't know how we got onto it* {pause} *er sh- # and I think she said something about oh she knew her tables and erm* {pause} *you know she'd come from Hampshire apparently and she #* {pause} *an- # an- yo- # you know er we got talking about ma- and she's taken her child away from* {pause} *the local school* {pause} *and sen- # is now going to a little private school up* {pause} *the Teign valley near Teigngrace apparently fra-*

It is hardly necessary to underline the differences between these data-samples, in terms of complexity and general "messiness". True, I have made the contrast particularly vivid by drawing my real-life examples from informal, spontaneous speech. But even written language which has undergone the disciplines of publication tends in practice to be much less predictable than the Pisa examples. Here are a few sentences drawn at random from the LOB Corpus of published British English:

> *Sing slightly flat.*
>
> *Mr. Baring, who whispered and wore* pince-nez, *was seventy if he was a day.*
>
> *Advice - Concentrate on the present.*
>
> *Say the power-drill makers, 75 per cent of major breakdowns can be traced to neglect of the carbon-brush gear.*
>
> *But he remained a stranger in a strange land.*

In the first example we find a word in the form of an adjective, *flat*, functioning as an adverb. In the next example, the phrase *Mr. Baring* contains a word ending in full stop followed by a word beginning with a capital which, exceptionally, do not mark a sentence boundary. The third "sentence" links an isolated noun with an imperative construction in a logic that is difficult to pin down. In *Say the power-drill makers...*, verb precedes subject for no very clear reason. The last example is as straightforward as the examples from the Pisa meeting; but straightforward examples are not the norm.

Not surprisingly, then, the task of annotating the grammatical structure of real-life language samples turned out to be a good deal more complicated than I had anticipated.

It would be inappropriate, here, to give much detail about the structural annotation scheme we evolved. For the sake of concreteness, I shall briefly illustrate it with an extract from the CHRISTINE Corpus (see FIG. 2.1). In FIG. 2.1, the next to rightmost field contains the words of a speech turn uttered by the speaker whose CHRISTINE code name is Gemma006, and the rightmost field shows the tree structure in which the words occur, displayed on successive lines as segments of a labelled bracketing. Gemma's second word *you* is a noun phrase (`N`) functioning as subject (`:s`) of its clause, whose verb group (`V`) is the single word *want*. The object of *want* is an infinitival clause (`Ti:o`), whose understood logical subject is again *you*, hence a "ghost" element `s101` is inserted in the word stream with an index number, 101, which marks it as identical to the subject of the main clause – and so on. The field to the left of the words contains their "wordtags" – word classifications drawn from an alphabet of 350-odd structurally-distinct classes of word; Gemma's first word, *well*, is classified as one type of "discourse item" (a form, characteristic of speech, which does not usually constitute part of a wider grammatical structure).

```
T02_0120    -----   Gemma006
T02_0123    .....   00328
T02_0126    0051229 *   UW      well    .
T02_0129    0051241 |   PPY     you     [S[Ny:s101.Ny:s101]
T02_0132    0051252 |   VV0v    want    [V.V]
T02_0135    0000000 y   YG      s101    [Ti:o.
T02_0138    0051264 |   TO      to      [Vi.
T02_0141    0051274 |   VV0v    nip     .Vi]
T02_0144    0051285 |   RP      over    [R:q.
T02_0147    0051285 |   RLh     there   .R:q]
T02_0150    0051303 |   CC      and     [Ti+.
T02_0153    0000000 y   YG      s101    .
T02_0156    0051315 |   VV0v    see     [V.V]
T02_0159    0051326 |   DDQ     what    [Fn?:o[Dq:G103.Dq:G103]
T02_0162    0051338 |   PPHS2   they    [Nap:s.Nap:s]
T02_0165    0051350 |   VV0i    come    [V.V]
T02_0168    0051362 |   II      on      [P:p.
T02_0171    0000000 y   YG      103     .P:p]
T02_0174    0051372 |   II      on      [P:p.
T02_0177    0051382 |   AT      the     [Ns.
T02_0180    0051394 .   NN1c    roll    .Ns]P:p]Fn?:o]Ti+]Ti:o]S]
```

Figure 2.1. Extract from CHRISTINE corpus

2. BUILDING TREEBANKS

When I began drawing trees for Geoffrey Leech's project, he produced a 25-page typescript listing a set of grammatical category symbols which he suggested we use, with notes on how to apply them to debatable cases. I

remember thinking that this seemed to cover every possible linguistic eventuality, so that all I needed to do was to apply Leech's guidelines more or less mechanically to a series of examples. I soon learned differently. Every second or third sentence seemed to present some new analytic problem, not covered in the existing body of guidelines. So I and the research team I was working with began making new decisions and cumulating the precedents we set into an ever-growing body of analytic rules. What grew out of a 25-page typescript was published in 1995 as a book of 500 large-format pages, *English for the Computer* (Sampson 1995). In the same year, an independently-developed but very comparable annotation scheme used for the University of Pennsylvania treebank was published via the Web, as Bies et al. (1995).

This great growth in annotation guidelines was caused partly by the fact that real-life language contains many significant items that are scarcely noticed by traditional linguistics. Personal names are multi-word phrases with their own characteristic internal structure, and so are such things as addresses, or references to weights, measures, and money sums; we need consistent rules for annotating the structures of all these forms, but they are too culture-bound to be paid much attention by the inherited school grammar tradition. In written language, punctuation marks are very significant structurally and must be fitted into parse trees in some predictable way, but syntactic analysis within theoretical linguistics ignored punctuation completely.

The more important factor underlying the complexity of our annotation rules, though, was the need to provide an explicit, predictable annotation for every turn of phrase that occurs in the language. As Jane Edwards has put it: "The single most important property of any data base for purposes of computer-assisted research is that similar instances be encoded in predictably similar ways" (Edwards 1992: 139).

For the theoretical linguists who set much of the tone of computational linguistics up till the 1980s, this kind of comprehensive explicitness was not a priority. Syntactic theorists commonly debated alternative analyses for a limited number of "core" constructions which were seen as having special theoretical importance, trying to establish which analysis of some construction is "psychologically real" for native speakers of the language in question. They saw no reason to take a view on the analysis of the many other constructions which happen not to be topics of theoretical controversy (and, because they invented their examples, they could leave most of those other constructions out of view). Language engineering based on real-life usage, on the other hand, cannot pick and choose the aspects of language structure on which it focuses – it has to deal with everything that comes along. For us the aim is not to ascertain what structural analysis corresponds to the way language is organized in speakers' minds – we have no way of knowing that; we just need some reliable, practical way of registering the full range of data in a consistent manner.

Often, so far as I can see, various different analyses of some usage would each be perfectly reasonable; our task is not to ask which analysis is "right", but to choose one of the analyses (at random, if necessary) and to make explicit that this is the analysis we have chosen, so that future examples of the same construction will be annotated in the same way, and statistics extracted from a treebank will not make the mistake of adding apples and pears. Consider, for example, the construction exemplified in *the more, the merrier* – the construction that translates into German with *je* and *desto*. Here are three ways of grouping a sentence using that construction into constituents:

[[the wider the wheelbase is], [the more satisfactory is the performance]]

[[the wider the wheelbase is], the more satisfactory is the performance]

[[[the wider the wheelbase is], the more satisfactory] is the performance]

The two clauses might be seen as co-ordinated (as in the first line), since both have the form of main clauses and neither of them contains an explicit subordinating element. Or the second clause might be seen as the main clause, with the first as an adverbial clause adjunct. Or the first clause might be seen as a modifier of the adjectival predicate within the second clause. There seems to be no strong reason to choose one of these analyses rather than another; what matters, if we are to produce meaningful statistics for use in language engineering, is to settle on one of the analyses and stick to it.

Theoretical linguists are sometimes inclined to look down their noses at this kind of taxonomic exercise as having little intellectual, scientific substance. Linguists of the Chomskyan, generative school have in the past been quite dismissive of taxonomic approaches[4]. But I do not see how the theoreticians can hope to make real progress in their own work, without a solid foundation of grammatical taxonomy to catalogue and classify the data which their theories ought to explain. In the comparable domain of natural history, it was two centuries after the taxonomic work of John Ray, and a century and a half after that of Linnaeus, before theoretical biology was able to develop as a substantial discipline in its own right in the late nineteenth century[5]. From a theoretical point of view the Linnaean system was somewhat "unnatural" (and was known from the start to be so), but it provided a practical, usable conspectus of an immensely complex world of data; without it, theoretical biology could not have got off the ground. The generative linguists' idea that they can move straight to theorizing, bypassing the painstaking work of taxonomy altogether, is one that I find baffling. In any case it is clear that language engineering, for which theories about psycholinguistic mechanisms are irrelevant, badly needs a comprehensive groundwork of data collection and classification.

One explanation for some linguists' lack of interest in the taxonomic approach may be their belief that natural-language structure is much simpler than it superficially appears to be. Underlying the diversity of observable linguis-

tic "performance", it is suggested, there lurks a mental grammatical "competence" which is the proper object of scientific linguistic study, and which can be defined in terms of a limited number of rules, to a large extent genetically inherited and common to all natural languages. As "competence" systems, the grammars of natural languages such as English or Czech might be hardly more complex than those of programming languages such as Pascal or Java[6].

If that were so, one might well see little need for intensive data-systematizing activity. But the idea that a natural language is governed by a limited set of clear-cut grammatical rules does not survive the experience of structurally annotating real-life examples. One way I sometimes express this point is to say that if someone cares to identify a rule of English grammar which they regard as reliable, I would expect to be able to find in English corpus data an example that breaks the rule, not as a momentary error of "performance" but within wording that appears to be intended by the writer or speaker and which "works" as an act of communication.

Thus, one rule which grammarians of English might take to be as reliable as any is the rule that reflexive pronouns, being co-referential with earlier elements in the same clause, can never occur as clause subjects. Yet here is an example found in the LOB Corpus, taken originally from a published magazine article on the Cold War:

> *Each side proceeds on the assumption that itself loves peace, but the other side consists of warmongers.*

This use of *itself* is an absolutely direct violation of the rule; yet the sentence seems perfectly successful in making its point. In case anyone should suspect that the example was produced by a non-native speaker whose English was poor, the author was in fact Bertrand Russell, one of the leading English intellectuals of the twentieth century; and his very next sentence, not quoted here, has another example of the same construction, making it quite clear that this use of reflexive pronouns was no careless "performance error" but a studied effect.

The truth is that rules of natural-language grammar are not binding laws like the laws of physics or chemistry. They are flexible guidelines, to which users of a language tend to conform but which they are free to adapt or disregard. This makes the task of categorizing real-life usage much more challenging than it would be if natural languages were like programming languages with a few irregularities added, but it also means that the task cannot be bypassed via aprioristic theorizing.

3. EXPLOITING THE SUSANNE TREEBANK

Despite the low esteem in which theoretical linguists held taxonomic work, I soon found that even a small-scale English treebank yielded new scientific

findings, sometimes findings that contradicted conventional linguistic wisdom. For instance, introductory textbooks of linguistics very commonly suggest that the two most basic English sentence types are the types "subject – transitive-verb – object", and "subject – intransitive-verb". Here, for instance, are the examples quoted by Victoria Fromkin and Robert Rodman in *An Introduction to Language* to illustrate the two first and simplest structures diagrammed in their section on sentence structure (Fromkin & Rodman 1983: 207-9):

the child found the puppy

the lazy child slept

Looking at statistics on clause structure in the treebank I developed at Lancaster, though, I found that this is misleading (Sampson 1987a: 90). "Subject – transitive verb – object" is a common sentence type, but sentences of the form "subject – intransitive-verb" are strikingly infrequent in English. If the sentence has no noun phrase object to follow the verb, it almost always includes some other constituent, for instance an adverbial element or a clause complement, in post-verb position. *The lazy child slept* may be acceptable in English, but it could be called a "basic" type of English sentence only in some very un-obvious sense of the word "basic".

Going a little deeper into technicalities, I was able to use the SUSANNE treebank to shed new light on an idea that has been known to most linguists since it was first discussed by Victor Yngve in 1960. Yngve argued that there is a constraint on the incidence of "left branching" in English parse trees[7]. Suppose, for instance, that we accept FIG. 2.2 (after Yngve 1960: 462) as an appropriate constituency diagram for the complex noun phrase *as good a young man for the job as you will ever find:* it is noticeable that the branches stretch down from the root much further in the "south-easterly" than the "south-westerly" direction. Yngve measured the left-branching factor of individual words numerically: for instance, *young* in FIG. 2.2 would be assigned the number three, because three of the four branches linking that word to the root are other than the rightmost branch below their respective mother nodes; and he suggested that there might be some fixed maximum, perhaps imposed by our mental language-processing machinery, such that grammatical structures are allowed to "grow" up to that maximum degree of left branching but not beyond it (whereas right branching is unlimited).

There is no doubt that Yngve was right to identify an asymmetry in English grammatical structure: the broadly N.W.-to-S.E. trend of parse trees is very noticeable in any English treebank. An investigation of the SUSANNE treebank, though, showed that the precise generalization was rather different from what Yngve supposed (Sampson 1997b). English does not have an absolute limit on the global left-branchingness of parse trees as wholes. It has a local and

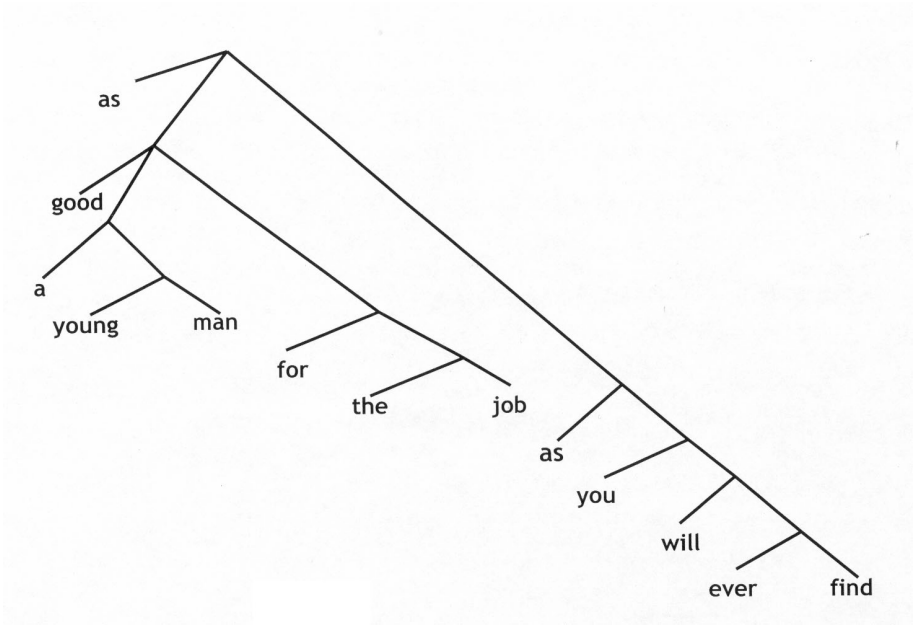

Figure 2.2. Tree for *"as good a young man for the job as you will ever find"*

statistical constraint on the left-branchingness of individual tree nodes: that is, there is a fixed probability for the expansion of any nonterminal symbol to contain another nonterminal symbol as other than the rightmost daughter, and, because this probability is low, in practice trees do not "grow" very far in the south-westerly direction. This was a rather pleasing finding: the fact that the constraint is local rather than global represents good news for computational tractability. Although Yngve's hypothesis had been a standard part of linguists' education for more than thirty years, there was no way of checking the facts before treebanks became available.

The statistics on left branching reveal a surprisingly precise kind of order underlying the diversity of natural-language constructions. Another kind of statistical investigation carried out on my original Lancaster treebank suggested a measure of precision in the extent of that diversity itself.

Richard Sharman (former Director of SRI Cambridge) has likened natural languages to fractal objects, which continue to display new detail no matter at what scale they are examined. An investigation I carried out on the frequencies of constructions in the Lancaster treebank made that analogy seem

rather exact[8]. I took a high-frequency syntactic category, the noun phrase, and counted the frequencies of the various expansions of that category in the treebank – for instance, the commonest expansion of the category "noun phrase" (in terms of the coarse vocabulary of grammatical categories which I used for this purpose) was "determiner + singular noun", which accounted for about 14% of all noun phrases in the data. But there were more than 700 other noun-phrase types which occurred at lower frequencies – many types each occurred just once in the treebank. It turned out that there was a regular relationship between construction-frequencies, and the number of different constructions occurring at a given frequency. As one looks at lower and lower frequencies, more and more different constructions occur at those frequencies, with the consequence that quite large proportions of text include constructions which are individually rare. Specifically:

> if m is the frequency of the commonest single construction (in my data, m was about 28 per thousand words)
>
> and f is the relative frequency of some construction (fm is its absolute frequency)
>
> then the proportion of all construction-tokens which represent construction- types of relative frequency less than or equal to f is about $f^{0.4}$

This finding contradicts the picture of a natural language as containing a limited number of "competent" grammatical structures, which in practice are surrounded by a penumbra of more or less random, one-off "performance errors". If that picture were correct, one would expect to find construction frequencies distributed bimodally, with competent constructions occurring at reasonably high frequencies, individual performance errors each occurring at a low frequency, and not much in between. My data were not like that; constructions were distributed smoothly along the scale of frequencies, with no striking gaps[9].

Consider what the mathematical relationship I have quoted would mean, if it continued to hold true in much larger data sets that I was in a position to investigate. (I cannot know that it does, but as far as they went my data contained no suggestion that the relationship broke down at the lower frequencies.) If the relationship held true in larger bodies of data, then:

- one in every ten construction-tokens would represent a type which occurs at most once per 10,000 words
- one in every hundred construction-tokens would represent a type which occurs at most once per 3.5 million words

- one in every thousand construction-tokens would represent a type which occurs at most once per billion words

One per cent of all construction-tokens is surely far too high a proportion of language material for language engineering to wash its hands of as too unusual to deal with. One might feel that one in a thousand is still too high a proportion for that treatment. Yet, if we have to search millions, or even hundreds of millions, of words of text to find individual examples, we seem to be a long way away from the picture of natural language grammars as limited sets of rules. A fractal object can never be fully described – one has to be willing to ignore detail below some cut-off. But the degree of detail which computational linguistics ought to be taking into account extends well beyond the range of structures described in standard grammars of English.

These lessons which computational linguistics is beginning to learn about the need to come to grips with the tremendous detailed complexity of real-life data, and the inadequacy of approaches based on aprioristic ideas of language structure (such as the ideas which led to the invented example sentences I quoted from the Pisa conference), are parallel to lessons which the wider information-technology world learned twenty to thirty years earlier.

4. SMALL IS BEAUTIFUL

When increasing computer power first made it possible to attempt large-scale software projects in the late 1960s, software development hit a crisis (see e.g. Sommerville 1992: 3, Ghezzi et al. 1991: 4). Repeatedly, projects failed or massively overran their time and resource budgets, because the individualistic, "craftsman"-like style of programming which had worked well for small projects, where all the coding might be done by one person, could not be made to scale up for large industrial projects which are necessarily executed by teams, and which need to be robust under many variations of conditions over long periods. The supervisor of a programmer executing a one-man project could see the lines of code he wrote, but could not see the many factors the programmer was tacitly taking into consideration in deciding how to write the code (and hence did not grasp their significance); and those many factors were typically far fewer than *should* have been taken into consideration, if the software was to have a long life – with one-man programs it is cheaper to throw them away and write new ones if conditions change, but with large-scale projects one cannot do that. Large-scale software development requires all the real-life complications which software may face to be identified and made explicit among members of a team, before a line of code is written.

The IT response to the software crisis of the years around 1970 was to create the discipline of "software engineering", which nowadays trains computing professionals to resist the instinctive view that coding is their primary activity.

Software engineering teaches entrants to that profession to devote the bulk of their efforts to identifying and spelling out on paper, in increasing detail, the requirements and specifications for the software to be produced. Coding the software is the trivial stage at the end of the process; if the specifications have been made sufficiently explicit, translating them into code is an almost mechanical process.

The computational linguists of twenty years ago who developed language processing software on the basis of unnaturally simple invented examples were like the one-man programmers of the early 1960s who produced software that crashed when extraneous circumstances changed. Since about 1990, computational linguists have come to understand that invented examples cannot do justice to the reality of natural languages, and that there is no substitute for corpus data. The discipline has, one might say, learned part of the software engineering lesson. But it has not yet learned the whole. The discipline, as I see it, has not yet grasped that availability of real-life data is not enough: one must put time and effort into studying the detailed structure of the data, before one can develop language-processing systems which are responsive to that structure.

This partial vision is apparent in the widespread emphasis on size of language corpora. In the pre-software-engineering days, IT managers tended to evaluate their programmers in terms of lines of code written per week: the ultimate purpose of the job was to produce program code, so it seemed that someone who produced code faster must be a better worker. Nowadays, in that domain we understand that premature or too-rapid code production is not a virtue. But in corpus linguistics, whether we are talking about "raw", unannotated corpora or about structurally- annotated treebanks, too many people still seem to feel that the first or only desideratum is size[10].

Other things being equal, obviously more data are better than fewer data. If a corpus were entirely devoid of analytic mark-up, size might be the only consideration – though, now that (for the principal Western European languages, at least) virtually unlimited quantities of written material are freely available via the Internet, constructing a large but completely raw corpus would be an exercise of questionable value. Even relatively "raw" standard corpora such as Brown and LOB involved a great deal of effort put into ensuring that they were balanced and properly representative of a well-defined genre; this effort mattered, and to some extent it militated against large size (though the main size constraint on those early corpora derived from the technology of the time, which required the samples to be manually keyed in).

When we come to structurally annotated corpora, however, the picture is entirely different. The whole point of a treebank lies in the detail, consistency, and accuracy of its structural annotations, and, perhaps even more important, in the extent to which the scheme governing the annotations has been ren-

dered explicit. All these things can be achieved only by methods which remain largely manual, so quality of annotation is a desirable feature which directly conflicts with quantity of annotated material. These features are both desirable, obviously; but one must be traded off against the other. At present, I sense a danger that computational linguists are much more aware of the quantity than of the quality factor (perhaps because quantity is more objective and easier to "sell" to research sponsors).

In the work of my group I have chosen the opposite priority: we treat detail, accuracy, and explicitness of annotation as more important than quantity of material annotated, with the inevitable consequence that our treebanks have to be relatively small. (The book which defines the SUSANNE annotation scheme contains quite a lot more wording than the SUSANNE treebank.) All agree that what is ultimately needed are very large quantities of very accurately annotated data. On the basis of the software engineering precedent, I would argue that we have a better chance of moving towards that ideal by taking care of analytic detail and accuracy first and worrying about quantity later, than the other way round. But alternative strategies may also be worth trying; my main point is that one should not be so dazzled by the quantity factor as to lose sight of the quality factor altogether.

5. ANNOTATING A SPOKEN CORPUS

Despite its small size, the SUSANNE treebank seems to have established itself over the seven years since its first release as a useful international research resource. I quite often encounter research based on it that was carried out by groups at distant sites of whom I had no prior knowledge. Now that my group has moved on, within the CHRISTINE project, to drawing trees for spoken language, many new analytic problems have been presenting themselves which again require detailed consideration, in order to develop consistent and logical annotation guidelines, and which in consequence again militate against corpus size.

How, for instance, are we to indicate what is going on in a "speech repair" – a passage in which a speaker corrects himself or changes tack in mid-utterance? FIG. 2.3 is an outline version (using spelled-out rather than coded grammatical category labels) of the CHRISTINE analysis of part of an utterance in which a speaker embarks on a relative clause modifying *any bonus* and then decides instead to use *anything* as the head of the phrase and to make *bonus* the predicate.

We use the # symbol to indicate the point where the speaker interrupts himself; but we need rules for deciding how to fit that symbol, and the words before and after it, into a coherent structure – do we, for instance, label what precedes the interruption point as a relative clause, even though only its first word, *that*,

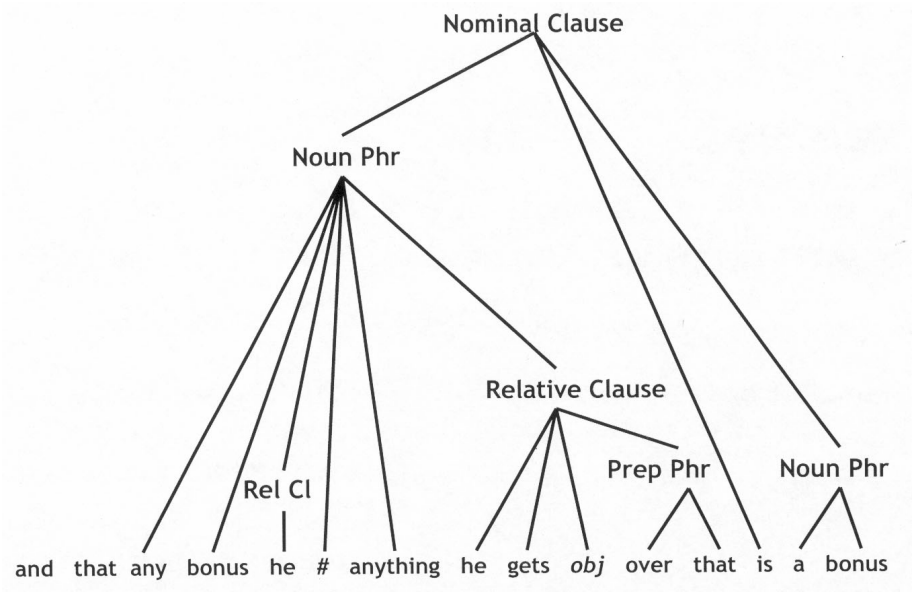

Figure 2.3. Tree from CHRISTINE Corpus

was actually uttered? Where in the tree do we attach the interruption symbol? The tree in FIG. 2.3 is based on explicit decisions about these and related questions, and the variety of speech management phenomena found in real-life spontaneous speech is such that these guidelines have had to grow quite complex; but only by virtue of them can thousands of individual speech repairs be annotated in a predictable, consistent fashion.

Structural annotation of spontaneous speech calls into question grammatical distinctions which, in writing, are fundamental. Written English takes pains to leave readers in no doubt about the distinction between direct and indirect speech, which is marked by punctuation even where the wording itself does not make the status of a particular quotation clear. Speech has no inverted commas, but commonly wording shows whether quotations are directly reported or paraphrased. For instance, in the following CHRISTINE example:

> *he said he hates drama because the teacher takes no notice, he said one week Stuart was hitting me with a stick and the teacher just said calm down you boys*

the words *he hates drama* (rather than *I hate ...*) show that the object of the first *he said* is indirect speech, whereas *hitting me* (rather than *hitting him*), and the imperative and second-person form in the quotation attributed to the teacher, show that the object of the second *he said* is a direct quotation which itself contains an internal direct quotation. But matters are not always so straightforward. Two more CHRISTINE examples run:

I said well that's his hard luck

well Billy, Billy says well take that and then he'll come back

The discourse item *well* at the beginning of *well that's his hard luck* usually marks the beginning of a direct quotation, but in context that would call for *your hard luck* rather than *his...* Again, following *Billy* the word *well* and the imperative *take* suggest direct speech, *he'll come* in place of *I'll come* suggests indirect speech. In spoken English it seems that directness of quotation is not an absolute property but a matter of gradience. Quotations may be reported more less directly, which creates a new challenge for an annotation scheme that was developed originally in connexion with written prose.

Indeed, the structures found in spontaneous speech sometimes call into question not merely the inherited range of grammatical-category distinctions but the very concept of grouping utterances into tree-shaped structures. It is no surprise to find that spontaneous utterances are sometimes too chaotic for any structure to be clearly assignable. More troubling are cases where the structure does seem clear, but it conflicts with the hierarchical assumption which is implicit in the use of tree diagrams or labelled bracketing. This applies to what I call "Markovian" utterances (which occur repeatedly in the CHRISTINE data), where a window of limited size moved through the wording would at each point contain a coherent, normal grammatical sequence, but the utterance as a whole cannot be assigned a single structure. Consider for instance the following, said by Anthony Wedgwood Benn, M.P., on a BBC radio programme:

> *and what is happening* {pause} *in Britain today* {pause} *is ay- demand for an entirely new foreign policy quite different from the Cold War policy* {pause} *is emerging from the Left*

The long noun phrase *an entirely new foreign policy quite different from the Cold War policy* seems to function, with respect to the words before it, as the complement of a prepositional phrase introduced by *for* which postmodifies *a demand*; yet, with respect to what follows it, the same phrase-token functions as subject of *is emerging*. Within the usual framework of grammatical analysis, one constituent cannot fill both of these roles simultaneously. Yet is it reasonable to abandon the concept of hierarchical grammatical structuring,

which has been entrenched in Western linguistic thought for centuries, because of a limited number of "Markovian" examples which seem to conflict with it?

I have argued elsewhere (Rahman & Sampson 1999) that some of the difficulties in evolving well-defined guidelines for consistent structural annotation of spoken language may stem from the fact that our inherited grammatical tradition has evolved almost exclusively in connexion with written language. It may be that fully-adequate schemes for annotating speech treebanks will eventually need to adopt notational devices that depart further from traditional grammatical ideas that anything yet adopted in our CHRISTINE Corpus. These are problems with which we are only beginning to grapple.

6. USING THE CHRISTINE CORPUS

Already, though, our data on structures in spontaneous English speech are beginning to yield novel scientific findings that simply could not be addressed before speech treebanks were available. I used an early version of the CHRISTINE treebank, containing most but not all of the material now circulated as CHRISTINE/Stage I, to examine correlations between grammatical complexity and speakers' demographic characteristics (Sampson 2001, chapter 5). By "grammatical complexity" here I mean what traditional school grammar means when it calls some sentences "simple" and others "complex": a complex sentences is one that contains a subordinate clause. Informal speech cannot be divided straightforwardly into "sentences", but one can ask, for each successive word of a speaker's utterance, how many clause nodes occur on the path between that word and the root note of the largest structure within which the word occurs; that gives a complexity figure for individual words, and a speaker can be assigned a figure for the mean complexity of all his words.

The only earlier research on this issue of which I am aware is that of Basil Bernstein (e.g. 1971: chapters 5-6), who claimed to find a strong correlation between grammatical complexity and social class: middle-class speakers produced utterances of greater complexity than working-class speakers. (This was one of the research results on which Bernstein based his famous theory of "restricted" v. "elaborated" linguistic codes.) However, Bernstein's analysis was based on a small data-set which was gathered for the purpose in a rather artificial manner: it by no means represented a fair sample of spontaneous usage – no such samples were available to Bernstein.

My analysis of the CHRISTINE material – which, although it has its own shortcomings, can claim to be a genuine cross-section of spontaneous speech in the United Kingdom in the 1990s – did not show correlations between grammatical complexity and social class (though I would not want to rest much reliance on a negative finding – there may be such correlations which failed to show up in our material). However, the data did contain positive evidence

of a different kind of grammatical/social relationship. There was a statistically significant correlation of grammatical complexity with age.

Of course, anyone would expect very young children to produce simpler utterances, on average, than mature speakers; but those who believe, with the generative linguists, in a "critical age" for language acquisition (e.g. Lenneberg 1967: 142ff., Pinker 1994: 290ff.) should predict that complexity will increase over the period until puberty, but will then flatten out as the innate biological clock closes the language acquisition machinery down and a speaker enters the period of his life during which he is a language-user but not a language-learner. Our figures were not like that. Complexity grew more slowly in adulthood than in childhood, but it continued to grow throughout life, so that, say, people in their fifties were producing slightly more complex utterances, on average, than people in their forties. This finding is, so far as I know, entirely novel, and is surely one of some human significance. (Having been born during the Second World War myself, I find it rather heartwarming.) It was only made possible by the creation of a speech treebank.

My new LUCY project, which will sample the writing of unskilled young adults as well as skilled producers of "model" documentation, will enable new kinds of question to be examined objectively for the first time, to my knowledge. Is the fumbling, imperfect writing of unskilled youngsters at the end of secondary schooling located (with respect to linguistic structure) on a direct route between the structure of spontaneous speech and that of model written prose, or does it have characteristics of its own, different from either of the latter? Is learning to be a skilled writer merely a matter of editing out features which are appropriate only to speech, or does it involve mastering structural techniques which are absent in spoken language?

People have ideas about these questions, but even the ideas of educated professionals in this domain are often based on guesswork which is sometimes demonstrably incorrect. That does not always mean that people overlook differences between spoken and written structure; sometimes, they see differences which do not exist. In the opening chapter of an influential book on this issue, Cheshire & Milroy (1993: 78) quote, as their paradigm example of a construction which is pervasive in spoken English but "avoided" or "unacceptab[le]" in the written language, what they call "dislocated" sentences, such as *These cats, they're hungry*. In reality, treebank data show that this construction is found in careful written English. Half the examples quoted when the construction is defined in the SUSANNE annotation scheme (Sampson 1995: §4.525) are taken from the highly literary genre called *belles lettres*. To date we know very little about structural differences between speech and writing, in the sense that our assertions are founded on objective, quantitative data.

7. CONCLUSION

We live surrounded by oceans of language; yet at the end of the twentieth century there are still vast areas of knowledge about language and its workings to be explored, which can only be approached through the painstaking work of developing structurally-annotated corpora of language samples. For those working on languages other than English, the existing resources are often less adequate than those with which we English linguists are blessed; but, by the same token, the range of virgin territory available for them to explore is all the greater. As a new century dawns, the future for treebank research looks bright indeed.

Notes

1. I believe the term "treebank", now in standard international use in this sense, was first coined by Geoffrey Leech in connexion with the project described here.

2. My home page, http://www.grsampson.net/, contains references to and locations of the treebank resources mentioned above. For the corpora mentioned below as providing language samples used in various treebanks, see http://nora.hd.uib.no/icame.html (Brown and LOB Corpora) and http://info.ox.ac.uk/bnc/ (British National Corpus).

3. I make no apology for repeating here examples which I used in order to make similar points in Sampson (1987a).

4. See, for instance, Chomsky's use of the term "taxonomic model" (1964: 11), or J.J. Katz's comments (1971) on linguistics as what he called "library science".

5. See, for instance, chapter 9 of Allen (1976).

6. For the doctrine of "competence" v. "performance", see Chomsky (1965: 4, 9, 15). Linguistic nativism is propounded in books such as *Reflections on Language* (Chomsky 1976) and *The Language Instinct* (Pinker 1994). I have pointed out the vacuousness of Chomsky's and Pinker's various arguments for linguistic nativism in my EDUCATING EVE (Sampson 1997a).

7. Others later suggested that the nature of the constraint may vary among different natural languages.

8. The analytic findings are presented in Sampson (1987b); see Sampson (1992: 440-5) for discussion of a critique of my conclusions by Briscoe and others.

9. It would in any case be surprising to find that "performance error" was an important factor in this treebank, which was based on published writing.

10. I have even heard a Swedish lady linguist comment, rather wickedly, that she realized that corpus linguistics was a male-dominated profession when she found that everyone with a new corpus wanted to tell her how big his was.

References

Allen, D.E. (1976). *The Naturalist in Britain: A Social History*. Penguin.

Bernstein, B. (1971). *Class, Codes and Control, vol. 1: Theoretical Studies Towards a Sociology of Language*. Routledge and Kegan Paul.

Bies, Ann, et al. (1995). *Bracketing Guidelines for Treebank II Style*. `http://www.cis.upenn.edu/~treebank/home.html`

Cheshire, Jenny, & J. Milroy (1993). Syntactic variation in non-standard dialects. In J. & Lesley Milroy, eds., *Real English: The Grammar of English Dialects in the British Isles.* Longman.

Chomsky, A.N. (1964). *Current Issues in Linguistic Theory*. Mouton.
Chomsky, A.N. (1965). *Aspects of the Theory of Syntax*. MIT Press (Cambridge, Mass.).
Chomsky, A.N. (1976). *Reflections on Language*. Temple Smith.
Edwards, Jane (1992). Design principles in the transcription of spoken discourse. In Svartvik 1992.
Fromkin, Victoria, & R. Rodman (1983). *An Introduction to Language* (3rd. ed.). Holt, Rinehart and Winston.
Ghezzi, C., et al. (1991). *Fundamentals of Software Engineering*. Prentice-Hall.
Katz, J.J. (1971). *The Underlying Reality of Language and Its Philosophical Import*. Harper & Row.
Lenneberg, E.H. (1967). *Biological Foundations of Language*. Wiley.
Pinker, S. (1994). *The Language Instinct: The New Science of Language and Mind*. William Morrow (New York). My page references are to the Penguin ed., 1995.
Rahman, Anna, & G.R. Sampson (1999). Extending grammar annotation standards to spontaneous speech. In J. Kirk, ed., *Corpora Galore*. Rodopi (Amsterdam).
Sampson, G.R. (1987a). Probabilistic models of analysis. In R.G. Garside et al., eds., *The Computational Analysis of English: A Corpus-Based Approach*. Longman.
Sampson, G.R. (1987b). Evidence against the 'grammatical'/'ungrammatical' distinction. In W. Meijs, ed., *Corpus Linguistics and Beyond*, Rodopi (Amsterdam). Reprinted in Sampson 2001.
Sampson, G.R. (1992). Probabilistic parsing. In Svartvik 1992.
Sampson, G.R. (1995). *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press (Oxford).
Sampson, G.R. (1997a). *Educating Eve*. Cassell. Revised paperback ed., Cassell, 1999.
Sampson, G.R. (1997b). Depth in English grammar. *Journal of Linguistics* 33.131-51.
Sampson, G.R. (2001). *Empirical Linguistics*. Continuum.
Sommerville, I. (1992). *Software Engineering* (4th ed.). Addison-Wesley (Wokingham, Berks.).
Svartvik, J., ed. (1992). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Mouton de Gruyter (Berlin).
Yngve, V. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104.444-66.