THE SYNTACTIC REGULARITY OF ENGLISH NOUN PHRASES

Lita Taylor, Claire Grover, Ted Briscoe[1]
Department of Linguistics
University of Lancaster
Bailrigg
Lancs., LA1 4YT, UK.

ABSTRACT

Approximately, 10,000 naturally occurring noun phrases taken from the LOB corpus were used firstly, to evaluate the NP component of the Alvey ANLT grammar (Grover et al., 1987, 1989) and secondly, to retest Sampson's (1987a) claim that this data provide evidence for the lack of a clear-cut distinction between grammatical and 'deviant' examples. The examples were sorted and classified on the basis of the lexical and syntactic analysis undertaken as part of the LOB corpus project (Sampson, 1987b). Tokens of each resulting type were parsed using the ANLT grammar and the results analysed to determine the success rate of the parses and the generality of the rules employed.

INTRODUCTION

In this paper, we present the results of an analysis of just over 10,000 English noun phrases (NPs) extracted from the Lancaster Oslo/Bergen (LOB) corpus treebank (Sampson, 1987b), a syntactically analysed 50,000 word subset of the 1 million word LOB corpus. The motivation for this research is twofold. Firstly, we wish to use this substantial data-base of naturally occurring constructions to test the accuracy and adequacy of a (purportedly) wide-coverage sentence grammar (Grover et al., 1987, 1989) which has been developed over the past three years as part of a general-purpose morphological and syntactic analyser for English (hereafter the Alvey Natural Language Tools (ANLT) grammar).[2] The research reported here forms part of an ongoing project to evaluate the complete grammar using data extracted from the LOB corpus (see Briscoe et al., 1987a). Secondly, Sampson (1987a) has analysed a large subset of the same NPs and argued that they provide evidence against any clear-cut distinction between grammatical and 'deviant' sentences in natural language. Sampson suggests that the lack of such a distinction precludes the possibility of successful automated natural language processing (NLP) using a generative grammar. If correct, this conclusion would have profound implications for our own work and the majority of other work in NLP (since the ANLT grammar is a type of generative grammar). Therefore, we wished to assess the evidence which Sampson uses to support his conclusion.

The LOB treebank is a manually analysed set of sentences drawn from the lexically analysed and tagged LOB corpus.[3] An analysis consists of a labelled bracketing containing lexical syntactic tags and phrasal or clausal 'hypertags'. Sampson (1987a:221) reports that there are 47 tags and hypertags relevant to the analysis of NPs – 28 lexical tags, 14 hypertags and 5 punctuation tags. Analyses are assigned to sentences according to the intuitions of the linguist guided by a 'casebook' of precedents (Sampson, 1987b). One important feature of these analyses is that the resulting tree structures are quite 'shallow' in the sense that there are rarely intervening nodes between the topmost node marked NP and the lexical tags themselves. Whilst most NP postmodifiers are treated as independent constituents, NP premodifiers are largely analysed as immediate daughters of the topmost NP node. In addition, punctuation tags are usually attached as immediate daughters of this node. A second significant feature of the LOB treebank analysis scheme is that tags and hypertags are atomic symbols (albeit with mnemonic names designed to indicate aspects of their featural composition).

Sampson (1987a:221) treats these 47 tags and hypertags as defining the types of distinct NP: "two or more noun phrases are regarded as tokens of the same type if their respective immediate constituents (ICs) represent the same sequence of possibilities drawn from this 47-member set of constituent-types". The example he gives of an NP type is DT* *S , F which would be the analysis assigned to an NP consisting of a determiner, plural noun, comma and finite clause. In this example, Sampson has generalised across sets of atomic tags through the use of 'wildcard' symbols, so DT* generalises across DTI, DT$, DTS, DTX, and so forth. He does not explain the extent to which he has generalised types in this fashion; however, since (hyper)tags contain at most four letters representing distinct features there are strict limits on featural decomposition within this framework of analysis. Sampson found that the 8328 NP tokens in his sample fell into 747 distinct NP types (relative to the notion of type just described). However, the crucial point of his argument is that the distribution of tokens amongst types is very wide. Sampson finds that there are a few very common types (such as 1135 tokens of DT* N* ie. determiner followed by noun) and a large number of distinct types with very few tokens (such as 468 types represented by a single token). Sampson examines the shape of the constituent type/token curve which results from analysing each type frequency relative to the most frequent type in the corpus. Sampson (1987a:225) concludes that this analysis provides "no evidence at all of a two-way partition of noun phrase types into a group of high-frequency, well-formed constructions and a group of unique or rare 'deviant' constructions; instead noun phrase types in the sample appear to be scattered continuously across the frequency spectrum." Furthermore, he suggests that the evidence from NPs supports his claim that "the range of constructions occurring in authentic texts seems so endlessly diverse

that the enterprise of formulating watertight generative grammars appears doomed to failure" (1987b:219).

The last step in Sampson's argument from the distribution of tokens amongst NP types to the failure of the generative paradigm is not made completely explicit. However, we believe that a legitimate way of reconstructing it is as follows. Suppose that we convert each NP type as defined above into a phrase-structure rule of a generative grammar (so DT* *S , F becomes NP –> DT* *S , F and so forth). Now consider the form that such a grammar will take: there will be a small number of quite general rules which will be used frequently and a very large number of particular rules used very infrequently. Crucially, for any corpus considered, many of the particular rules will be motivated by just one token in the data. Thus, these rules are not rules in any genuine sense since they express no generalisations over the data. Furthermore, this suggests that the task of the generative linguist (in search of watertight grammars) will never be complete because each new set of data will bring with it the need for further highly idiosyncratic 'rules' of this kind.

Whilst it seems likely that "all grammars leak" slightly, one clear problem with Sampson's argument is that his evidence only bears on one particular and implausible generative grammar, rather than on the paradigm as a whole. It may well be that the generalisations which can be expressed in terms of a phrase-structure grammar employing a finite set of (nearly) atomic categories are not those appropriate to elegant description of natural language syntax (Chomsky, 1957; Gazdar et al., 1985). In addition, the strategy of adopting 'shallow' analyses in which each phrase-structure rule will have many daughter categories will tend to reduce the applicability of each rule. In these respects, the ANLT grammar is a more conventional generative grammar, based on recent monostratal approaches to syntactic description. Syntactic categories are feature complexes and unification is employed as the method of grammatical combination. Syntactic generalisations are expressed in terms of partially specified immediate dominance rules, linear precedence rules and a variety of metagrammatical statements concerning feature defaults, propagation, optional pre/postmodification, and so forth.[4] In addition, the particular analysis of NPs adopted recognises a number of intermediate nominal categories (such as N-bar), as well as recursion within these categories, and this ensures that most individual rules mention fewer daughters than would be typical in the analysis used in the description of the LOB treebank. For these reasons, we felt that a fairer test of Sampson's claims would be to evaluate the same corpus of NPs with respect to the ANLT grammar. In addition, this excercise would provide valuable information concerning the real adequacy of the account of English NPs incorporated into this grammar.

THE ANALYSIS TECHNIQUE

A superset of the corpus of data analysed by Sampson (1987a) was extracted from the LOB treebank using tree searching software developed by the first author and Roger Garside of Lancaster University's computing department. Following Sampson, we ignored categories G (Belles lettres, biography, essays) and P (Romance and love story) from the treebank data-base. The omission of this treebank data merely reflects the state of development of the treebank at the time when Sampson undertook his experiment. However, Sampson also ignored coordination because he felt that coordination reduction and such phenomena would create "special complications". We include results for the coordinated examples because the ANLT grammar contains the required rules. In other respects, the initial samples are identical; both being drawn from an identical 38,212 word sample from the treebank.

Of the 10,150 NPs in this sample of the treebank, 17 were rejected because they were incorrectly analysed and either were not, in fact, NPs or else the boundaries of the putative NP were incorrectly marked and, therefore, our access software failed. The remaining 10,133 NPs were initially sorted into single and multi constituent NPs (according to the LOB model of analysis). Single constituent NPs were further sorted according to the incidence and order of their immediate lexical constituents and multi constituent NPs according to the incidence, order and attachment of their immediate daughters. At this point, we discarded a further 119 NPs which were tagged in a way which indicated they contained either foreign phrases (for example, *fait accompli*) or mathematical formulae and symbols. These are tagged but not analysed internally in the treebank. We assume that they are irrelevant to the syntax of English NPs. These steps resulted in 10,014 NPs being sorted into 2358 distinct NP types. These types must be identical with Sampson's initial analysis (modulo the inclusion of coordination and exclusion of formulae and foreign phrases) because they are based entirely on the literal form of the tags in the LOB treebank.

The next stage of our analysis was to semi-automatically reduce these 2358 NP types into fewer types by collapsing together tags on the basis of grammatical generalisations exploited in the ANLT grammar rules and implicit in the LOB tag names. For example, there is no purpose in treating NPs identical apart from the number of the head noun as distinct (although they are tagged distinctly) because the ANLT grammar will deploy precisely the same set of rules to analyse them. Sampson (1987a) also collapsed types by generalising across tags, however, he gives no details of this procedure, so it is impossible to quantify the extent to which our analyses diverged at this point. Following Sampson, we ignored the internal structure of post-modifiers (such as PPs, relative clauses, etc.) and of possessive premodifiers. However, in order not to trivialise the experiment we analysed the same set of lexical data covered by his analysis *regardless of whether lexical items are treated as immediate constituents of NP in the ANLT grammar*. For example,

sequences of simple adjectival or possessive premodifiers are directly attached to the topmost NP node in the treebank, so we consider these cases in our results.

We also performed some manual editing of the LOB examples to remove punctuation. The ANLT grammar contains no rules referring to punctuation since we do not regard punctuation as a syntactic phenomenon. However, where punctuation reflects a genuine syntactic distinction (such as that between restrictive and non-restrictive postmodification), examples were classified appropriately. This approach probably gives us a slight edge over Sampson in terms of the generalising power of our rules, but we do not regard this as pernicious because we do not recognise a syntactic difference between examples such as *the man with red shoes in the park* and *the man with red shoes, in the park*, given the semantically intuitive analysis. 48 NPs contained brackets, of which 34 signalled appositional or parenthetical material. The appositional cases were parsed with brackets deleted. The parenthetical cases were counted as failures (see below for further discussion). In 8 of the remaining cases, the brackets were internal to an embedded constituent and were, therefore, irrelevant. 3 further examples contained point numbering or marking (i.e. a)... b)...) conventions and the final 3 enclosed ordinary modifiers. These 6 examples were parsed with brackets and numbering/marking conventions removed.

These steps resulted in 707 distinct NP types. Sampson (1987a) found 747 types. When one considers that punctuation will have increased the number of types he found, it seems likely that we have probably reanalysed the data in a manner quite similar to his original analysis. One token of each of the 707 revised types of NP was parsed using the ANLT grammar NP rules. Initially, we attempted to perform this analysis automatically using the ANLT project parser in batch mode. The words in the example to be parsed were replaced with their lexical tags and a 'lexicon' was created relating tags to lexical syntactic categories in the ANLT grammar. Data from the treebank and other data from two different corpora were parsed in this fashion and the output was manually analysed to select the semantically correct analysis, weed out 'false positives' where the system had assigned one or more incorrect analyses, and to diagnose the reasons for parse failure.

Failures occurred both because of inadequacies in grammatical coverage and because of resource limitations with some long and multiply-ambiguous NPs. The resulting data contained many cases of multiple analyses of the type expected using a grammar containing rules to handle PP attachment and compounding (see, for example, Church & Patil, 1982). The intention was to compute the frequency with which each rule of the grammar applied and the overall success rate of the grammar/parser from these manually edited files. However, the process of evaluating and searching for correct analyses amongst very high numbers of automatically generated parses required more effort than manually applying the rules to check that the semantically correct analysis could be produced. This problem highlights the need for automatic semantic 'filtering' of the parses produced, but, in the absence of a fairly comprehensive and sophisticated lexical and compositional semantic component, this was not possible.

Therefore, we completed the analysis of one token of each of the 707 NP types by manually applying the ANLT grammar to check that the semantically appropriate analysis could be produced. When the correct parse was available, the rules used in this analysis were recorded. We derived a numerical index of the generality of each rule by counting each application and multiplying it by the number of tokens in each type exemplified by the parsed example.

RESULTS

622 of the 707 examples were parsed successfully, yielding a success rate of 87.97% When the success rate takes account of the frequency of each NP type in the sample and indicates the proportion of successful NP parses which would be achieved by the ANLT system for this data, the figure rises to 96.88% or 9702 NPs parsed successfully out of the 10,014 sample.

The analyses utilised a total of 54 distinct rules expressed in the ANLT 'object grammar' formalism. Of these 8 were additions prompted by the experiment: 3 for names (*Mr. Joe Bloggs*), 1 for noun compounding (*water meter*), 2 for adverbial pre- and post-modification (*nearly a century*), 1 for possessive NPs dominated by N-bar (*the America's cup*), and 1 for NPs with adjectival heads (*the poor*). We added these rules because they express uncontroversial generalisations and represent 'oversights' in the development of the grammar rather than ad hoc additions solely for the purposes of the experiment.

These object grammar rules were produced by 7 linear precedence statements, 4 rules of feature propagation, 6 feature default rules, 3 metarules, and 50 immediate dominance rules in the metagrammar. Although the metagrammar is the 'seat of linguistic generalisations' in our system, parsing proceeds in terms of a compiled object grammar derived from these metagrammatical statements. Therefore, statistics concerning rule application will be associated with the object grammar.

We counted the number of times each of the 54 object grammar phrase-structure rules would apply in the analysis of all the parsable examples in the sample. The categories of these object grammar rules still contain features with variable-values which will be instantiated at parse time by unification. They are therefore considerably more general than similar rules with atomic or nearly-atomic categories (of the kind which are implicit in the treebank analyses and resulting NP types). Table 1 below presents these results. The rules used and their corresponding names are a superset of those described in Grover et al. (1987). Grover et al. (1989) describes in detail all the rules used below.

Table 1 – Number of Applications of the 54 Object Grammar Rules

Rule Name	No. of Applics.	Brief Explanation
CONJ/N1A	141	N1 conjunct, no coordinator
CONJ/N1B	133	N1 conjunct, with coordinator
CONJ/N2A	423	N2 conjunct, no coordinator
CONJ/N2B	382	N2 conjunct, with coordinator
CONJ/NA	14	N conjunct, no coordinator
CONJ/NB	13	N conjunct, with coordinator
N/COORD1	12	*and* coordination of N
N/COORD2A	1	*or* coordination of N, all conjuncts with same PLU value
N1/COORD1	43	*and* coordination of N1
N1/COORD2A	57	*or* coordination of N1, all conjuncts PLU –
N1/COORD2D	33	*or* coordination of N1, all conjuncts PLU +
N2/COORD1A	358	*and* coordination of N2
N2/COORD1B	7	*and* coordination of N2 but no coordinators (i.e. a list)
N2/COORD2	2	*both..and* coordination of N2
N2/COORD3A	17	*or* coordination of N2, all conjuncts PLU –
N2/COORD3C	1	*or* coordination of N2, differing PLU values
N2/COORD3D	1	*or* coordination of N2, all conjuncts PLU +
N/ADJ	159	N –> ADJ – *the poor* and adjs. in compounds
N/COMPOUND	1054	N –> N N – *water meter*
N/NAME1	127	Names – *Tom Brown, A. N. Other*
N/NAME2	206	Names with pre- and post-titles – *Mr. Brown, J. Brown esq.*
N/NAME3	3	Complex titles – *vice president, prime minister*
N1/APMOD1	2134	Prenominal AP modifier
N1/APMOD2	190	(2 versions to restrict number of attachments)
N1/INFMOD	2	Infinitival VP postmodifier with gap – *the man to ask*
N1/POSS	13	The possessive morpheme *'s*
N1/POSSMOD	3	Possessive NP as premodifier – *the America's cup*
N1/POST_APMOD	43	AP postmodifier – *the man most likely to win*
N1/VPMOD	184	Passive or progressive VP postmodifier – *the man dying/killed*
N1/PPMOD	777	PP postmodifier
N1/REL	352	Relative clause postmodifier
N1/N	7170	An N with no complements
N1/PP	1132	PP complement
N1/SFIN	2	Sentential complement
N1/VPINF	6	Infinitival VP complement
N2+/DET	4534	N2[+Spec] –> DET N2[-Spec] – *the book*
N2+/PART1	7	Partitive, plural – *many of the books*
N2+/PART1(FOOT6)	1	Wh version – *how many of the books*
N2+/PART2	86	Without *of* – *all the books*
N2+/PART3	20	Partitive, singular – *each of the books*
N2+/POSSNP	146	Possessive NP in specifier position – *the man's book*
N2+/PRO	1974	Pronouns
N2+/PRO(FOOT9)	1	Wh pronouns
N2+/PRO2	111	Pronouns in partitives
N2+/QUA	185	Quantifying adj. in specifier position – *all books*
N2-	7819	N2 with no specifier – *books*
N2-/QUA	380	Quantifying adj. in non-spec. position – *(the) many/three books*
N2-/QUA(FOOT4)	1	Wh version – *how many books*
N2/ADVP/1	47	Adverbial phrase premodification
N2/ADVP/2	32	Adverbial phrase postmodification
N2/APPOS	274	N2 –> N2 X2[+Prd] – apposition/non-restrictive modification
N2/COMPAR_1	8	Comparative NP with *than* PP – *more books than him*
N2/NEG	10	N2 –> *not* N2
POSSNP	12	Possessive NP – *the man's*

There are a number of reasons why some of these figures are slightly misleading. For example, some low numbers are an artifact of the preliminary analysis into types. Thus, N2+/PRO(FOOT9), which would be utilised to parse NPs consisting of wh-pronouns, such as *who*, *what*, and so forth, only applies once. In the preliminary analysis, we decided to collapse together tags for the wh and non-wh version of the same category. It is just an accident that in all of the representative tokens of each type which were parsed, only one wh-pronoun turned up and this happened to represent a singleton type. Similarly, N1/SFIN only applies twice, but it is probable that there are more examples of nouns taking sentential complements as arguments in the sample. The LOB tagset represents these complements by 'Fn' and relative clauses by 'Fr'. Following Sampson, we collapsed all of these to 'F'. Consequently, the bulk of the sentential complements were incorrectly added to the types involving postmodification by relative clauses. These problems are unavoidable, given the particular assumptions built into the LOB treebank analyses, unless a completely new analysis of the sample was undertaken.

One way of ameliorating this problem is to collapse some of the distinct rules in Table 1. A number of the distinct object grammar rules are present for 'technical' reasons connected with the use of fixed-arity unification and feature propagation by variable binding in the ANLT grammar formalism and parser (see Briscoe et al., 1987b,c for details). Therefore, we reduced the 54 object grammar rules to 36 hypothetical rules using our judgement to determine whether a distinction between rules was motivated by a linguistic generalisation or a technical consideration peculiar to the ANLT grammar formalism. In most cases, the linguistic generalisation is, in fact, present in the metagrammar rules but 'compiled out' in the automatic production of the equivalent object grammar. For example, rules with 'FOOT' in their name are wh-variants of other rules defined by metarules which state the manner in which they differ (systematically) from the non-wh versions. The resulting 36 hypothetical rules are given in Table 2 along with new rule application counts based on summing the counts for the merged actual rules. We also give the figures for the number of times each rule applied in the parsing of one token of each type. The final column presents a 'proportioned-up' figure based on multiplying the second column by 15.6 (since the parsed tokens represent 6.41% of the total sample). This column gives another perspective on the 'generalising power' of the rules involved.

Table 2 – Applications of 36 Hypothetical Rules

Rule Name	Total No. of Applics.	No. in Parsed Tokens	Proptiond.-up Total
CONJ/N1	174	18	281
CONJ/N2	805	106	1654
CONJ/N	27	17	265
N1/COORD	133	8	125
N2/COORD	389	42	655
N/COORD	13	8	125
N/ADJ	159	28	437
N/COMPOUND	1054	216	3367
N/NAME1	127	34	530
N/NAME2	206	47	733
N/NAME3	3	3	47
N1/APMOD	2324	288	4493
N1/INFMOD	2	2	31
N1/N	7170	598	9329
N1/POSS	13	9	140
N1/POSSMOD	3	3	47
N1/POST_APMOD	43	22	343
N1/PP	1132	67	1045
N1/PPMOD	777	144	2246
N1/REL	352	70	1092
N1/SFIN	2	2	31
N1/VPINF	6	4	62
N1/VPMOD	184	45	702
N2+/DET	4534	320	4992
N2+/PART	114	26	406
N2+/POSSNP	146	38	593
N2+/PRO	1975	29	452
N2+/PRO2	111	24	374
N2+/QUA	185	36	562
N2-	7819	552	8611
N2-/QUA	381	92	1435
N2/ADVP	79	37	577
N2/APPOS	274	157	2449
N2/COMPAR_1	8	6	94
N2/NEG	10	7	109
POSSNP	12	8	125

COMPARISON OF RULES AND TYPES

We suggested above that Sampson's argument against the generative concept of grammaticality is based on the assumption that each type in his original analysis will be associated with one rule. Sampson (1978a) found 747 types of which 468 were singleton types containing only one token, or 62.65% singleton types. In our reconstruction of Sampson's analysis we found 707 types of which 421 were singleton types, or 59.95% singleton types. Sampson's commonest type contained 1135 tokens, ours contained 1519 tokens. Sampson (1987a) presents an analysis of his data which involves plotting a frequency-ordered list of NP types against the cumulative frequency of NP tokens in types of the same or lower frequency. This allows him to predict that 'rare' types, defined in terms of rate of occurrence relative to the rate of occurrence of the commonest type, will crop up fairly often in naturally occurring samples of NPs. For instance, if 'rare' is defined as occurring no more than once per 1000 occurrences of the commonest type, then about one example in 16 will represent some rare type. Therefore, a robust parser will need many 'rules' for such 'rare' types. Furthermore, there is no reason to expect the percentage of singleton types to fall as the sample size grows, implying that a robust parser of unrestricted text deploying a finite set of generative rules is out of the question.

Unfortunately, we cannot repeat Sampson's analysis for both our types and our rules because more than one rule is involved in the parsing of many of the types. Using the ANLT NP rules, an average of 5 rules applied

to each parsed token exemplifying a type, this figure drops to 3.18 when we take the average for the complete sample. Therefore, there is no direct correlation between rules and types. Nevertheless, Sampson's result follows directly from the high proportion of singleton types in his analysis and his assumption that one rule will suffice for each type; as he writes "although a rare type is by definition represented by fewer tokens in a sample than a common type, as we move to lower type-frequencies the **number of types possessing those frequencies grows**, so that the total proportion of tokens representing all "rare" types remains significantly large even when the threshold of "rarity" is set at relatively extreme values." (Sampson, 1987:225, original emphasis).

The most basic and important difference between any grammar based on a one-to-one correspondence of rules and types and one such as the ANLT grammar is the enormous difference in its size; namely, 36 or 54 rules as opposed to 707 or 747 rules – reduction by a fac-tor between 13 and 20 approximately. This alone testifies to the greater generality of the ANLT NP grammar rules. However, there are also big differences in the patterns of application of rules between the two approaches. We can see this by looking at an ordered list of the rarest 10 types and comparing it with similar lists for the least applied actual and hypothetical 10 ANLT rules. The first column in Table 3 shows the number of tokens or rule applications. Following columns show numbers and percentages of types or rules associated with this number of tokens or applications.

Table 3 – 10 Least Frequent Types / -ly Applied Rules

No. of Toks./ Rule Applics.	Number of Types	Number of Actual Rules	Number of Hypthetcl. Rs.
1	421 (60%)	6 (11%)	0 (0%)
2	84 (12%)	3 (6%)	2 (6%)
3	46 (7%)	2 (4%)	0 (0%)
4	21 (3%)	0 (0%)	0 (0%)
5	16 (2%)	0 (0%)	1 (3%)
6	12 (2%)	1 (2%)	1 (3%)
7	3 (.5%)	2 (4%)	0 (0%)
8	7 (1%)	1 (2%)	1 (3%)
9	8 (1%)	0 (0%)	0 (0%)
10	5 (1%)	1 (2%)	1 (3%)
12	–	2 (4%)	1 (3%)
13	–	2 (4%)	2 (6%)
14	–	1 (2%)	0 (0%)
27	–	–	1 (3%)
43	–	–	1 (3%)
79	–	–	1 (3%)
111	–	–	1 (3%)

Summing the percentage values reveals that 88.92% of tokens fell into the ten rarest types, 38.89% of actual rules fell into the ten least applied classes, and 33.33% of hypothetical rules fell into the ten least applied classes for that set. Table 3 further demonstrates the greater generality of the rule-based analysis versus the type-based analysis for this sample of NPs. But in a sense, presenting the results in this manner misses the crux of Sampson's argument that any parsing system based on generative rules will need a large or open-ended set of spurious 'rules' which simply redescribe the data, because they will only apply once. In the actual rule set, 6 rules or 11.11% are dubious in this sense, but, as we argued above, these rules are only distinct for technical reasons and in the hypothetical set no such rules exist. In any case, the proportion of actual dubious rules represents a considerable improvement on the proportion of singleton types (59.55%).

In (1) we present 3 (randomly-chosen) tokens of NPs from singleton types. If Sampson's general thesis were correct, we would expect such examples to be exotic or syntactically mysterious.

(1)
a) the old tension-bar-sprung Morris Minor
b) the main existing indirect tax , purchase tax
c) a basic ideological one

These NPs are not problematic for the ANLT grammar and are classified as singleton types because of the nature of the lexical and syntactic analysis used in the LOB treebank. Similarly, ANLT rules which applied 'rarely', such as N1/VPINF (6 times) or N1/INFMOD (2 times), which would apply in the parsing of *desire to grow up* and *man to ask* respectively, do not encode controversial or doubtful generalisations. Although the actual frequency of such constructions in English may well be low.

THE FAILURES

It is instructive for similar reasons to examine those examples that the ANLT grammar failed to parse. If Sampson's general thesis were correct, we should expect these to fall into singleton types and be syntactically exotic or mysterious. In fact, they are relatively easy to classify and the failure of the ANLT grammar results from either intentional or in some cases unintentional 'oversights' in the NP grammar. The failures can be classified, as illustrated in Table 4.

Table 4 – Analysis of Failures

Classification	No. of Types	No. of Tokens
Odd Numbers	5	10
Dates	4	24
Ellipsis	11	122
Parentheticals	19	58
Right-node Raising	3	10
Odd Premodifiers	11	21
Paired Constructions	16	46
Unlike Category &	2	4
Miscellaneous	14	17

Odd numbers include examples like *2 Kings 25 : 25 , 6*, and so forth. No rule was included in the grammar for dates, although these all consist of day (written *10* or *10th*), month (unabbreviated), and year (in numerals). In 2 of the 4 cases the order of day and month is reversed. Ellipsis of the head noun in cases where there is a postmodifier, for example, *those who perpetuate it*, causes a problem for the ANLT grammar because the determiner *those* cannot be analysed as a pronoun since

the grammar blocks modification of pronouns. This problem accounts for all the failures in this class.

Parenthetical or intrusive material which is not in apposition comes in two kinds. Firstly, there are cases of grammatical modification which occurs between the head noun and its arguments, as (2) illustrates.

(2) our failure *over two centuries* to sustain any strong national musical tradition of our own

These are not parsed as a result of the rigid assumptions about the ordering of arguments and modifiers built into the grammar. These need to be relaxed on the basis of some theory of 'heaviness' and its effect on order. Secondly, there are cases of genuine intrusive interjection or interpolation, as (3) illustrates.

(3) little capsules , *this big , – he brandished a teaspoon* – with hundreds of tiny little red men inside them

Such intrusive material can occur in most positions from a syntactic perspective. We suspect that a theory concerning their distribution would be largely pragmatic.

Some cases of 'right-node raising' of phrases are covered by the ANLT grammar. However, there is no rule for 'right-node raising' of nouns which would appear to be needed in NPs such as *late 19th- and early 20th-century Rumania*. Similarly, the grammar restricts NP premodifiers to AP, but a number of non-AP premodifiers occurred in the sample. These mostly involved measure phrases of some form, such as *a 6 p.c tax free distribution*, *the 24ft passenger cabin*, or *the 5 shilling shares*. There are 4 cases of unlike category coordination in AP modifiers like *music , both manuscript and printed* and *wine-glass or flared heels*. The ANLT grammar allows this in post-copular position, but clearly the relevant generalisations should be extended to AP pre- and post-modifiers.

There are a number of cases where a premodifier selects a particular postmodifier. Comparative constructions with *more* and *than* are a well-known type which the ANLT grammar covers. However, there are many other more or less idiomatic phrases of this type, some of which could probably be subsumed by an expanded treatment of comparatives along existing lines, some of which could not. We give illustrative examples in (4).

(4)
such a crazy spin that Leslie could not cope with it
as much God's handiwork as a man
as little as 0.001 at % of the addition elements

In addition, the rule for noun compounding we have included does not allow compounds to contain anything other than lexical nouns. Cases of adjectives in compounds were treated as 'successes' by allowing the rule N/ADJ which converts adjectives such as *poor* to nouns to deal with ellipsis of the head noun in *the poor* to overapply to adjectives in compounds. In this area, the ANLT grammar is clearly inadequate and needs improvement in obvious directions. The rule N/ADJ should be replaced by a lexical rule which states that '+human' adjectives can function as nouns, and compounding rules should be allowed to cross the 'boundary' between morphology and syntax, perhaps by allowing N-bar categories as well as nouns to 'compound'. These modifications would allow the illustrative examples in (5) to be counted as successes.

(5)
the third geologists' association excursion
our well organised after care departments

The miscellaneous class contains 2 types where *each* occurs at the NP boundary, such as *silicon , copper and magnesium each*. We suspect that in these examples *each* should be treated as an adverbial modifier of the following VP. There are two types containing the phrase *all but* as part of a partitive, some cases of words, such as *no one* occurring unhyphened, and one or two more exotic examples illustrated in (6).

(6)
in *17 something* Newton discovered gravity
' a man on the roof ' by Kathleen Sully , Peter Davies, 15 shillings

A final example worthy of consideration is given in (7).

(7) the company's Caravelle schedules London-Brussels and onwards from Athens to various points...

This could be classified as a case of non-constituent coordination of NP and PP postnominally or as a case of specialised ellipsis of *from* before *London* in 'travel-agent-speak'.

CONCLUSION

Our results demonstrate quite clearly that a feature-based unification grammar employing a recursive and 'deeper' style of analysis captures the relevant generalisations more efficiently than the analysis and implicit formalism employed by Sampson (1987a). We have reduced approximately 700 types to between 36 or 54 grammatical generalisations about NPs and shown that a minimally modified generative grammar developed (largely) independently of the test corpus is capable of covering 96.88% of the sample considered. We can demonstrate concretely why this should be so by considering the distinct single-constituent NP types from the treebank data exemplified by DT* JJ N*, DT* JJ JJ N*, and so forth. In the ANLT grammar this potentially infinite set of types is analysed through the recursive application of four rules of the following broad type: NP –> DET N1, N1 –> AP N1, AP –> A, N1 –> N. Thus a potentially infinite set of NP types is reduced to 4 grammatical generalisations.

We do not wish to claim that we have developed a 'watertight' perfect grammar of the English NP (although we do feel that the ANLT grammar has withstood this evaluation very well). There is still the 3.12% or 312 NPs that we are unable, at present, to analyse, and there is good reason to believe that "all grammars leak" slightly. However, there is little evidence in our results to suggest that a few rule-governed grammatical generalisations about naturally occurring NPs of English

do not effectively demarcate grammatical examples; or to suggest that the enterprise of generative grammar is doomed because of the high proportion of rules required to deal with residual, particular cases. On the contrary, our analysis of the failures demonstrates that, for the most part, they are not parsed because of oversights in the ANLT grammar, rather than because they are deviant in syntactically mysterious ways.

Sampson (1987a:226) concludes that the "onus must surely be on those who believe in the possibility of NL analysis by means of comprehensive generative grammars to explain why they suppose that the shape of constituent type/token distribution curves will be markedly different from the shallow straight line suggested by our limited – but not insignificant – database." However, Sampson's result is suggested by his *analysis* of this data, not the data itself. In this paper, we have demonstrated that a more satisfactory analysis of essentially the same data-base leads to precisely the opposite conclusion.

In other respects, the conclusions we should draw from this experiment are less positive. The development of wide-coverage grammars for robust parsing of unrestricted text will only be achieved through extensive evaluation using naturally occurring data. This, in turn, rests on the availability of suitably structured corpora from which the relevant data can be extracted automatically and on suitable software for semi-automatically testing rules against this data. The ANLT batch-mode parsing system proved completely inadequate to the latter task (largely because it was developed to check the grammar against a hand constructed set of short illustrative, deliberately unambiguous examples). Sampson (1987a) was able to perform a more sophisticated analysis of the treebank sample precisely because the original structuring of the data corresponded to his 'theory of grammar and grammatical analysis'. The problems we have had making use of his analysis to preliminarily classify the same data in order to evaluate the ANLT NP grammar highlight the impossibility of developing a corpus databank structured in some grammatically 'descriptive' or 'uncontroversial' fashion (pace Sampson, 1987b).

FOOTNOTES

1. The first two authors are also members of and wholly funded by the speech and language research group IBM (UK) Scientific Centre, Athelston House, Winchester. The third is now at the Computer Laboratory, University of Cambridge, Corn Exchange St., Cambridge, CB2 3QG, UK.

2. The development of this analyser was funded by the Alvey Programme and involved three collaborating research projects at the universities of Cambridge, Edinburgh and Lancaster (Briscoe et al., 1987b; Phillips & Thompson, 1986; Russell et al. 1986).

3. See Johansson & Hofland (1987) for a description of the tagged LOB corpus and Leech et al. (1983) for a description of the lexical disambiguation and tagging procedure.

4. See Briscoe et al. (1987b) for a full description of the ANLT grammar formalism and Grover et al. (1987, 1989) for a description of the English grammar expressed in this formalism. Shieber (1986) provides an introduction to unification-based approaches to generative grammar.

REFERENCES

Briscoe, E.J., Craig, I. & Grover, C. 1987a. The use of the LOB corpus in the development of a phrase structure grammar of Emglish. In Meijs (1987).

Briscoe, E.J., Grover, C., Boguraev, B.K. & Carroll, J. 1987b. A formalism and environment for practical grammar development. *Proc. of IJCAI*, Milan, pp. 703-8.

Briscoe, E.J., Grover, C., Boguraev, B.K. & Carroll, J. 1987c. Feature defaults, propagation and reentrancy. In Klein, E. & van Bentham, J. eds. *Categories, Polymorphism and Unification.* Centre for Cognitive Science, University of Edinburgh, pp. 19-35.

Chomsky, N. 1957. *Syntactic Structures.* Mouton, The Hague.

Church, K. & Patil, R. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *Computational Linguistics*, 8, 3-4, 139-49.

Garside, R., Leech, G. & Sampson, G. 1987. eds., *The Computational Analysis of English: A Corpus-based Approach.* Longman, London.

Gazdar, G., Klein, E., Pullum, G.K. & Sag, I.A. 1985. *Generalized Phrase Structure Grammar.* Blackwell, Oxford.

Grover, C., Briscoe, E.J., Carroll, J. & Boguraev, B. 1987. The Alvey natural language tools grammar. *Lancaster Working Papers in Linguistics*, 47.

Grover, C., Briscoe, E.J., Carroll, J. & Boguraev, B. 1989. The ANLT grammar (2nd release). *Technical Report No. 162*, Computer Laboratory, Cambridge University.

Johansson, S. & Hofland, K. 1987. The tagged LOB corpus: description and analyses. In Meijs (1987).

Leech, G., Garside, R. & Atwell, E. 1983. The automatic grammatical tagging of the LOB corpus. *ICAME News*, 7, 13-33.

Meijs, W. 1987. ed., *Corpus Linguistics and Beyond.* Rodopi, Amsterdam.

Phillips, J.D. & Thompson, H.S. 1986. A parser for generalised phrase-structure grammars. *Edinburgh Working Papers in Cognitive Science*, 1, 115-137.

Russell, G.J., Pulman, S.G., Ritchie, G.D. & Black, A. 1986. A dictionary and morphological analyser for English. *Proc. of Coling86*, Bonn, pp. 277-279

Sampson, G. 1987a. Evidence against the "grammatical/ungrammatical" distinction. In Meijs (1987).

Sampson, G. 1987b. The grammatical database and parsing scheme. In Garside et al. (1987).

Shieber, S. 1986. *An Introduction to Unification-based Approaches to Grammar.* CSLI Lecture Notes 4, University of Chicago Press, Chicago.