The Present Use of Statistics in the Evaluation of NLP Parsers

J.Entwisle
Flinders University of South Australia
jim@cs.flinders.edu.au

D.M.W.Powers
Flinders University of South Australia
powers@cs.flinders.edu.au

Abstract

We are concerned that the quality of results produced by an NLP parser bears little, if any, relation to the percentage-results claimed by the various NLP parser-systems presently available for use. To illustrate this problem, we examine one readily available NLP tagging and parsing system, the ENGCG parser; and one tagger, the Brill tagger. We note responses to both artificially generated and naturally occurring text. The percentage assessments are methodologically flawed, and should be taken with a grain of salt; instead, assessment of the performance of an NLP parser should be effected by a user, and solely from a consideration of the resulting parses of exactly the input which an NLP user decides to contribute for such an assessment. Careful attention to input of whatever corpus the user decides on, is presently the only suitable qualifying test of parsing ability. The parsers available are none of them perfectible yet, despite apparent yields now quoted at 99%+. We consider the impact of Zipf's argument of 'least effort' on percentage assessment; and we open a discussion on estimating the relative complexities of corpora.

1 Introduction

Statistics are frequently bandied around in NLP, and would seem to be the obvious way to compare competing systems and methodologies. For example:

> As a rule, data-driven systems rely on statistical generalisations about short sequences of words or tags....[T]hey tend to reach a 95-97% accuracy [and 12 parsers are referred to.] Interestingly, no significant improvement beyond the 97% "barrier" by means of purely data-driven systems has been reported so far...[Then there is a report of three hybrids - systems that employ linguistic rules for solving some ambiguities - with various additions; and these hybrids] seem capable of exceeding the 97% barrier.... Next, a new system..uses only linguistic distributional rules. Tested against a 38,000-word corpus of previously unseen text, the tagger reaches a better accuracy than previous systems (over 99%). (Voutilainen, 1995)

We are concerned about the misleading nature of such published statistics, although researchers working on NLP systems are of course well aware that the figures must be interpreted carefully. To be able to show our cause for concern we must look at the statistics of a well known parser, and also at its actions. Unfortunately there are relatively few systems which are freely available for consultation or examination, and for this reason we are forced to pick on some of the few systems that are. The criticisms we make here are *not* directed against the parsers we use as examples, but against the way in which our field is treating its statistics.

Our main example is the ENGCG parser, in one specific form which has been available for several years now, and from which many influential variants have been spawned, including the one being used to tag the Bank of English (Voutilainen and Silvonen, 1996). Perhaps unfortunately for it, the ENGCG parser is very conveniently consulted on the Net, and statistics on it have been published. Another good reason for our examination of this parser-group is the comment made by them, above, that "the tagger reaches a better accuracy than previous systems (over 99%)." And the fact is that it does seem to represent the best approach currently.

2 Assessment of the ENGCG parser

The ENGCG configuration we are focussing on is actually a collection of tools or layers: morphological analyser, morphological disambiguator, POS tagger, finite-state parser and heuristic enhancement programs, etc. (Karlssen, 1990), (Tapanainen and

Jim Entwisle and David Powers (1998) **The Present Use of Statistics in the Evaluation of NLP Parsers.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 215-224.

Jarvinen, 1993), (Voutilainen, 1995), (Voutilainen, 1997), (Voutilainen and Silvonen, 1996).

In this paper, we will use the name "the ENGCG parser" as referring to a specific combination of these components in the manner that they are reported in Voutilainen (1995); in particular we are using the version which augments the underlying tagger with a finite-state parser, with heuristics switched on.

In considering and assessing the statistics on this parser, and also the parser itself, we record that we should not like it thought we are only trying to criticise that program. So let us repeat that in our view the ENGCG parser appears to be at least as effective as any other parser presently generally consultable or available, and that it may well be the most effective member of that class today. Indeed, we see that parser as defining the standards for a corpus-based parser. But, as will be noted, we nonetheless think the ENGCG parser parses poorly, when it is compared with human parsers - a conclusion we have reached partly because of matters discussed in §3.

2.1 The ENGCG parser - a statistical assessment by its creators

The creators of the ENGCG parser make it clear that their parser is not expected to parse contrived sentences, and they acknowledge that sentences can be created that will result in a less favourable response than they report; they also mention that their parser is intended to work on such things as manuals and they acknowledge that it works less well on fictional literature. Their claim and specifications are:

> 1,200 "grammar-based" constraints
> 99.7-100% of all words retain the appropriate morphological reading
> 3-7% of all words remain (partly) ambiguous
> 200 "heuristic" constraints
> resolves some 50% of remaining ambiguities after heuristic disambiguation, 99.5% or more retain the appropriate morphological reading
> ...The performance figures are measured against fairly neutral running English of the written variety. Similar performance on e.g. invented laboratory sentences is not guaranteed.
> (Voutilainen and Silvonen, 1996).

As to corpus, Voutilainen and Tapanainen (1993) refer to their tagging scheme as having "been manually applied on some 20,000 words of running text from various genres as well as on some 2,000 test sentences from a large grammar (Quirk *et al*, 1993), as a first approximation of the inventory of syntactic structures in written English..." This allowed them to validate their approach informally and to "ascertain the generality of the proposed rules" (Voutilainen and Tapanainen, 1996, p4). Voutilainen (1995, p9) later used "a 38,202-word test corpus consisting of previously unseen journalistic, scientific and manual texts" to test the ENGCG implementation of this scheme.

2.2 The ENGCG parser - our statistical assessment

Our own assessment of the operation of the ENGCG parser, over the first 1000 words of Chapter 3 of "Alice in Wonderland"(Carroll, 1975) (which might not be described as "fairly neutral"; but it is our own domain of study), is that the parser reported 108 (10.8%) words with extra wrong roles alongside the correct role, and 33 (3.3%) words with wrong roles and no accompanying correct role. 27 of the wrong roles that had no accompanying correct role were nouns that were mis-designated as to case, but were correctly designated as nouns. We did notice that there is no special tag specifically for the objects of non-finite verbs, and that these words were designated as having the label *nom* (i.e. "nominal") as also were the objects of prepositional phrases; this caused an extra 10 responses which we regarded as unusual, but which were not counted as incorrect because they seemed to be correctly designated according to the designers' specifications.

Further, there were a number of words with more than one role in positions where the words were clearly ambiguous; and these too were accepted as being correct. Thus, our results for the ENGCG parser on the first 1000 words of chapter three of "Alice in Wonderland" is that 89.2% of the words had no extra wrong roles and 96.7% of the words had a correct tag amongst their final list of roles.

We have conducted a further assessment on the operation of the ENGCG parser, over the first 1000 words of Chapter four of *Alice in Wonderland* (Carroll, 1975); the results there were less impressive - 82.0% of words had no extra wrong roles (18.0% did), and 94.8% of words had a correct tag amongst their final list of roles (5.2% did not).

Our conclusions regarding the ENGCG parser, and these figures on it, are that:

1 for what it does, it works well, and speedily too.

2 the difficulty that this parser faces is due mainly to the limitations on the starting roles (tags in the lexicon) on each word at the commencement

of parsing; the limitations are those supplied from a restrictive lexicon.

3 The ENGCG parser can only throw roles away, never gain any; so there is no way of redeeming any limitations inherited from the lexicon.

We see point 2 as the critical one. We suggest that a tagger or parser should not arbitrarily restrict the starting roles of a word, lest that restriction happen to exclude some legitimate parse. Probably the assumption in the ENGCG parser regarding starting-roles for words is not a feature that can easily be removed from this parser, but we will not speculate further on this matter except to note that reliance on this restricted set of roles pre-empts some parsing decisions where valid possibilities are not acknowledged. These roles are never added back in later, so their removal decreases the number of roles in the final results. An alternative approach to parsing, which makes no such lexical restriction on roles, is presented by Entwisle and Groves (1994), but exploring this further is beyond the scope of this paper.

3 The limitations of the ENGCG starting role lists

We now proceed to demonstrate ENCG's dependency on the starting-roles, which we note are decided upon partly from semantic considerations; for this purpose we create and submit to the ENGCG parser some artificially generated sentences.

Our sentences themselves might be considered to be contrived. Thus the parser's creators may say that our test sentences were not within the domain of their device. But keep in mind what we are demonstrating here: we are not arguing that the parser just fails, nor that it fails because it does not parse these reasonably average sentences. We are demonstrating that such failures reflect the high dependency of this parser on starting from this more presumptuous position; that such parses effected by this parser themselves demonstrate that that position is not an acceptable starting point for NLP.

We report parses submitted, below. Sentence (1): has four words that have each two starting roles (noun, verb); each role being fairly equally balanced in three of those words in English corpus. Sentence (2): has words similar to sentence (1) but one word instead is disproportionately balanced, so one role is much more likely. Sentences (1) and (2) are syntactically identical. Sentence (3) shows a parsing failure consequent on the limited starting-roles given to the word changed for sentence (2). Useful words for demonstrating these limitations are words like *watches* and *sails* and *fish*, all readily noun or verb.

In these parses, each indented line following a word in angle brackets has a separate role (or roles) for that word: e.g.*man* is given roles of subject and object; *fish* has three roles.

Our first test sentence is:

(1) The man who sails boats watches fish.

The ENGCG parser could say little about this; it did say:

"<*the>"
    "the" <*> DET CENTRAL ART SG/PL @DN>
"<man>"
    "man" N NOM SG @SUBJ @OBJ
"<who>"
    "who" <**CLB> PRON WH NOM SG/PL @SUBJ
"<sails>"
    "sail" V PRES SG3 VFIN @+FMAINV
"<boats>"
    "boat" N NOM PL @OBJ
"<watches>"
    "watch" N NOM PL @SUBJ @OBJ
"<fish>"
    "fish" <P/for> V INF @-FMAINV
    "fish" <P/for> V PRES -SG3 VFIN @+FMAINV
    "fish" N NOM SG/PL @OBJ
"<$.>"

Somehow *watches* has lost the role of main verb, which a native speaker would say was its only available role. And there is a second syntactic parse in that sentence, not reported by ENGCG. That parse is: *the:article man:subject [who:relative-pronoun sails:verb] boats:verb watches:indirect-object fish:object*. Such a parse is, of course, semantically most unlikely. Note that *boats* here is working as a ditransitive verb (Entwisle and Groves, 1994). The verb *ships* has this subcategorisation (as in *He ships them supplies each week*), more commonly than *boats*.

The ENGCG parser's reliance on starting-roles is clear, because of the different result when a sentence of the same inflection pattern, but with the word *water*, a word more commonly a noun, substituted for *fish*, is submitted:

(2) The man who sails boats watches water.

which gave:

"<*the>"
    "the" <*> DET CENTRAL ART SG/PL @DN>
"<man>"
    "man" N NOM SG @SUBJ @OBJ
"<who>"
    "who" <**CLB> PRON WH NOM SG/PL @SUBJ
"<sails>"
    "sail" V PRES SG3 VFIN @+FMAINV

“<boats>”
  “boat” N NOM PL @OBJ
“<watches>”
  “watch” <InfComp> V PRES SG3
    VFIN @+FMAINV
<water>”
  “water” N NOM SG @OBJ
“<$.>”

a better result; ENGCG provided a parse that traditional grammar requires, as well as a spurious parse; the reason for the improvement is clear when we examine that parser's response to a sentence where *water* has the less usual, but not unusual, role of verb, in:

(3) I water the plants.

“<*i>”
  “i” <*> PRON PERS NOM SG1 SUBJ @SUBJ
“<water>”
  “water” N NOM SG @NPHR
“<the>”
  “the” DET CENTRAL ART SG/PL @DN>
“<plants>”
  “plant” N NOM PL @NPHR
“<$.>”

The symbol *nphr* designates a “stray NP”. The possibility that *water* could be a verb is not entertained by this parser; and a single, completely incorrect parse is the only result: *noun noun determiner noun*. An impossible answer. Note too: ENGCG signals an improper parse: but fails to signal the failure[1].

We tested the ENGCG parser further: with other words of a similar nature, to show that this error was not a feature of just one word, *water*. We created a list of sentences to test starting-roles specifically:

(1) The man who sails boats *watches* fish.
(2) The man who sails boats *watches* water.
(3) I *water* the plants.
(4) Let him *water* the plants.
(5) The man *sands* wood.
(6) He *fords* streams.
(7) They *dog* his life.
(8) The *past* fades.
(9) He *watches* watches.
(10) They *dynamite* bridges.
(11) He *compliments* them.
(12) He *ferrets* out answers.
(13) He stopped *hunting* rabbits.
(14) He *sights* along the rifle.

[1]In this connection, the separate matter of permitting an “I don't know” response is an important feature in the correct approach to NLP but is not relevant in this paper.

The italicised word in each sentence is the test-word, the word of interest. Table 1 reports results.

The ENGCG parser is relying to some degree on limiting the starting roles of words to only the more likely ones, then it starts at that point to parse - but that restriction still allow multiplication of the parses that ENGCG offers - out to twelve, for the case of sentence (1). This parser then attempts only to give the more likely reading(s), but it does not necessarily offer a legal parse. Because of starting-role restrictions, this parsing program is not always producing an acceptable parse, which is unsatisfactory. Indeed, we view the two clear parsing failures regarding sentences (1) and (3), in the conditions we have established, as further evidence of a somewhat weak parsing action, one which fails to use all of the syntactic constraints in English, properly.

We ran these sentences in the more recent ENGCG-2 Tagger(Voutilainen 1995) and received parses that were improved, but the substantial point remained, although to a lesser extent: that tagger had not received the instruction to accept the verb *to water* as a fully equipped verb, and so a variation in sentence (3) - to

(4)*Let him water the plants.*

returned *water* as noun only. Nor did that tagger report the second parse in either of sentences (1) or (2).

Because the above sentences are artificially generated, the ENGCG parser has been working in a domain beyond its design; but this, the full domain of all written English, is our interest. So, the aim of the ENGCG parser is not that of unequivocally parsing English, in NLP.

The limited set of roles provided by the lexicon probably arises from omission, but we note two consequences arising from reduction of the set of possible tags through omission of valid, but rare roles. The first is obvious: if there is less disambiguation to perform, the tagger and parser will be faster. The second is perhaps less obvious: if rarer roles are omitted from a parser then it is incapable of correctly resolving them. It is possible to manipulate speed and error rate by judicious omission of roles.

We learn from the above demonstration that when a word of a sentence submitted to the ENGCG parser needs a lower-probability word-role, it may not find it; in fact it returned a strange “four nouns make a sentence” type of parse. It was not constrained by a grammar rule of English to say that *water* cannot be a verb - but by an arbitrary re-

System	*Alice* Chapter 3 - 1000 words		*Alice* Chapter 4 - 1000 words		Our test sentences
	wrong extra roles	missing wanted roles	wrong extra roles	missing wanted roles	test-word role missed
ENGCG parser	10.8%	3.3%	18%	5.2%	11/14
Brill tagger	4.9%	4.9%	5.1%	5.1%	12/14
Obvious tagger	13%	18.4%	15.5%	18.6%	10/14

Table 1: A comparison of three NLP programs

straint, taken from a starting-role list. In parsing English, the proper constraints must be used; no genuine rule of English arbitrarily limits a word's starting-role in this manner.

3.1 Consideration of the result

The foregoing may appear to have been very critical on one parser, and on just one point. So let us here remind you what was the aim of §3: to show that a parser can be covering over a lesser "parsing-action"; here, by a lexicon's restrictions on starting-roles for words. We had to document fully that precise claim to make clear our complaint. After we make a comparison with results from another tagger, we will make use of those findings in the subject of this paper - parser statistics.

4 The Brill tagger

The second NLP system that we will examine is the older Brill tagger; we will first report on it, and then compare results from both ENGCG parser and Brill tagger. The problem of limited starting-roles, indicated in §3 as causing parsing failures for the ENGCG parser, is seen in the Brill tagger also.

We found a slight complication in comparing the two devices; the Brill tagger offered only one tag per word. So, in applying the scoring system above, each incorrectly tagged word is firstly a wrong extra tag and then secondly a word without a correct tag - §5.1. This counting method might appear a little unfair to Brill: the error counts doubly. However, consider firstly that that scoring system follows the same scheme that was applied to the ENGCG parser previously (where we had no choice but to score it that way, because that program could, and did, have one or more correct tags, or a combination of correct and wrong tags, or wrong tags only - on a word.) Secondly, Brill said it was offering the alternative of more than one tag (but it just did not for our test text), thereby conceding the propriety of that scoring. Thirdly, we assess that scoring as proper, because English sometimes offers ambiguous syntactic roles on words; and an NLP parser must allow for actual English usages.

In Table 1 are the results of tests on both the ENGCG parser and the Brill tagger - in respect of our above list of 14 test sentences, and also of the two sections of *Alice in Wonderland* detailed in §2.2. The two separate results relating to each *Alice* chapter are "extra wrong roles" and "missed correct roles". The last column, titled "our test sentences", records the count of the number of mis-calculated roles, on each italicised word of our fourteen test sentences of §3 (scored at maximum of one error per sentence).

In particular, we note that those two parsing programs, ENGCG and Brill, failed to parse eleven and twelve words (respectively), out of those fourteen test-words contained in our testing sentences. We believe that our test sentences are not unnatural in any way.[2] Those results indicate to us that, in those programs, the starting-roles of most of those fourteen italicised words have been unduly limited. Indeed, the figures in the last column (titled "our test sentences") may have found a problem affecting all corpus-based parsers. Such a matter is not be relevant to this paper, and anyway we ourselves have not yet further assessed those figures, either.

As noted earlier above, we have found the ENGCG parser to be the most effective and informative of all readily-available current NLP parsers, and the strongest parsing program (though the results tabulated above do not completely vindicate our choice.) On this account, we have given above the detail that the ENGCG parser displays in its parses - for that information; and we will restrict comment and example hereafter, to the ENGCG parser.

5 Statistical measurement of parsers

We are now in a position to consider the matter of statistics and parsers. We discuss the current attempted statistical measurement of the quality of parsing programs. We find these measurements unconvincing, and indicative of a flawed methodology -

[2] We do consider sentence (9) to be a little unnatural: that sentence was submitted so as to test further on the word *watches*: and in the event, both programs found the correct parse there, so we feel bound to include it. We have reported all the sentences we tried.

for the reason that the measurements that have been adopted are not properly indicative of any gain-or-fall index in the ability of a parser to handle the syntax of the natural language of English.

We find that there are difficulties over the use of these statistics in the evaluation of the qualities of a parsing program. Take a concrete case for which in the foregoing section we have established the fact: that the ENGCG parser failed to have the starting-role of *verb* included for the word *water*. This failure, this omission, would almost certainly *reduce* the number of wrong roles that this parser would offer on almost any given corpus - unless the corpus just happened to include a substantial number of references to something about watering gardens, or similar expressions: that latter usage, of the verb "to water", whilst not unusual, is probably not going to be the usage met with by a parser using the type of corpus that the ENGCG parser advertises. Even with a number of occurrences of applying water, (i.e. using *water* as a verb), the parser will still probably show better results by omitting the possibility of it being a verb: for the most part the parser will operate "more accurately" from the omission, so long as the number of verb-*water* is fairly small.

There is no verb-role given to the word *water* by the ENGCG parser in its response to our example-sentence (2) above (that "non-response" was of course correct); and the noun role given by it is correct not only for sentence (2) but perhaps for about 99% of the appearances of the word *water*. *Water*, the noun, is a word very commonly met. So, the omission of the consideration of a correct role for the case of sentence (3) above, happens to have improved the statistics for the parse of sentence (2); and that is so, even though another result is that another ordinary sentence - sentence (3) - is wrongly parsed.

How can that deterioration in the ability of a parser to handle English generally, be considered to be an improvement in parsing quality? We see this as evidence of a methodology that demands simple improvement in some percentage-statistics for a reason which logically might have been considered instead as causing a decay in that percentage: the reasoning behind such a variation of those statistics appears to us as flawed. These particular statistics for parsers, then, are meaningless: the figures could, if one really wished it, easily be increased right up to 99% (or even higher) even if *water* appeared once or twice as a verb. That enhancement could be done by omitting every alternative word-role from the starting-role list if that alternative is only seldom ever needed in the given corpus: that would cause substantial statistical enhancement of a parser's reaction-measurements on even a very large, genuinely naturally occurring, corpus.

It could be argued that, because the parser then follows the main-chance reading of English, it is now actually the better parser for that. It is probably the faster parser, and it will respond with less spurious roles to many sentences, but it lacks depth in its deductions as a result; crucially in our view, that parser would have failed to tap any deep regularities of the patterns of English. Our own view is that until a parser is seen to be successfully handling sentences like sentence (3) without compromising its results on sentence (2), we do not care how good its statistics are on even randomly chosen corpora: such figures measure little, in our view. We do not suggest that starting-roles here have been tampered with at all, let alone deliberately sieved for statistical improvements: we say that absurdities such as increasing the positive statistics of a parser by reducing its parsing ability, and such opportunities for sieving as above, mean that those statistics are valueless for evidence of parser ability, let alone for comparison between parsers. Once we see that a parser is considering English with a fair degree of depth in its appraisal, we are prepared to consider that its response to randomly chosen corpora has meaning; but not until then. Such a use of statistics not only impairs the way parsers are evaluated but undermines the whole idea of statistical NLP.

5.1 Types of error in an NLP parser

There are two separate errors that we are looking at in relation to any computer-based parsing system; they both relate to the grammatical role or roles that a parser is offering in respect of a word. The first of these is: the failure to offer a correct role (i.e. a role is missing.) The second is: the offering of an extra, incorrect role. These two errors are essentially what is known in statistics (Speigel 1972) as, respectively, Type I and Type II errors - names which we will use hereafter. The two errors are also analogous to the errors of under- and over-generation respectively.

We say that, if a parser is both losing wanted roles (Type I error) and also gaining extra roles (Type II error), each error to a non-negligible degree, the quality of the parser cannot be measured; and so the figures for ENGCG parser, of 10.8% and 3.3% (and 18% and 5.2%) - §2.2 - are of further concern to us on this count.

If one were to make small changes to a parser, changes that did not fundamentally alter the parsing procedure but did alter the Type I - Type II balance, then the consequent reduction in one of those errors would result in an increase in the other. That much is clear. It is, however, unlikely that the two errors will move proportionately. On the contrary, for many systems, a change that causes a slight reduction in either one, causes a large increase in the other, since parsers are generally balanced at a point that sets both errors to a reasonable compromise. But this point is often a finely-balanced and sensitive, one. We point out this problem to make it clear that an error-pair of 18% and 5.2% respectively does not imply, say, a "total error" of 23.2% or an "average error" of 11.6% or any other such simplistic formulation. In fact, Samuelsson and Voutilainen (1997) show a hyperbolic relationship between the two errors (for the case of a 1988-style HMM-based parser). They name that relationship "the Error-Rate-Ambiguity Trade-Off".

Thus, we claim that if both kinds of error in a parser are non-negligible, neither figure means much. So, while "watches" and "water" (and other words) are being denied verb-roles (i.e. error Type I is non-negligible), the power of the parser in the other direction (error Type II) has little meaning. How bad will the "extra wrong roles error" (Type II) need to become before the "roles lost error" (Type I) vanishes? We cannot tell, from the information which those two last-mentioned writers provide. For ourselves then, we comment that only when we see that sentences like *He dogs my life* and *They dynamite bridges* are getting good consideration (so error Type I has become minimal), can an "extra wrong roles error" (error Type II) percentage mean something.

This is an alternative, and more general way, of viewing the concern that is the central point of this paper. By not permitting a parsing program to consider fully all the possible syntactic-roles on a word (by the use of limiting starting roles, say) error Type I is artificially raised somewhat, which lowers error Type II, possibly by a large amount; that forms our major concern.

We refer above to a suggestion of keeping one of the errors *minimal*. If possible, of course, that error should be brought to 0%, but the question of whether that is feasible is not going to be considered here.

Error - ambiguity

Voutilainen (1995) uses the term *error* for "error Type I" and *ambiguity* for any extra syntactic roles reported by his parser. We are concerned over such nomenclature because we cannot tell if these extra syntactic roles are wanted (i.e. genuine extra readings) or unwanted (i.e. errors of Type II). *Ambiguity* is not necessarily an error; several genuine readings may actually exist in the original text.

Genuine syntactic ambiguity is a completely different matter from any part of the subject of this paper; and this is so, even though there are two syntactic readings of each of sentences (1) and (2): we do acknowledge that this means that the two sentences are each *syntactically* ambiguous (but surely not semantically so!).

6 The choice of a corpus in an NLP parser

We find that some kinds of naturally occurring corpora never give a convincing display of the power of a parser, when the parser is tested on them alone - or indeed measured by reference to them alone. Some of what occurs in naturally occurring text is very standard and repetitive in structure: technical writing and newspaper reports use language that is very regular and repetitious in its patterning. Reams of this kind of text may have to be fed into a parser before that parser has been exposed to even a few of the many different syntactic phenomena that the language of English presents. And yet, all the time, these reams of text are prejudicing the statistics; the "percentage correct" count is nearing 100%, for a reason that is completely irrelevant to the parsing expertise of a parser - yet that, and only that, should be all that is being measured.

6.1 The corpus approach to parsing

We do accept that the use of large collections of naturally occurring text is at least partly a resolution of the problems inherent in small parsers of the past, in parsers that were shown to work only on one or two special phenomena. To that extent, we welcome the corpus-based approach as a response to such an offering. But as we have shown, the corpus-based approach is not without its own quirks. In summary: it is not the number of *words* that a parser can successfully parse (or even sentence statistics, which are dramatically lower). The number of different linguistic phenomena and the amount of syntactic diversity that the parser can successfully handle are the only proper measures of a parser's power to parse; that is what should be being measured by parser statistics.

We postulate that the more syntactically complex the corpus, the more a trial of the parser has been effected. In this regard researchers should welcome the appraisal of their parsers by others' use of invented sentences. But we would never defend the use of some absurd 'trap' of a sentence, contrived not for enquiry, clarification or proper test, but merely to score points against a parser. A charge of 'unnaturalness' in a sentence can be easily resolvable by native speakers or even by formal experiment, but only once the sentence has been placed in the appropriate context. A better option would be to find the required construct (or the actual sentence) in a corpus.

We believe that artificially generated natural sentences should be used freely by people other than the authors of the target in order to decide on the quality of the parsing program, but, as is common practice currently, a report by the creators should use corpus-generated measurements only.

We are unhappy with riders like - "Similar performance on e.g. invented laboratory sentences is not guaranteed" (Voutilainen and Silvonen, 1996): as being almost intimidatory to critics.

7 Estimating syntactic complexity of corpora

We consider some of the potential to variety in linguistic corpora, and we have also suggested some possible evaluations of that variety.

7.1 Application of Zipf's work

Zipf's concept of 'least effort' may be relevant here (Zipf, 1947; Powers, 1998): Zipf argues that the simpler a construct is, the more often humans will want to use it - or the more often they need it, the simpler they will make it. Zipf's concept applies to complex syntactic constructions too, and Zipf himself has demonstrated that his law applies even to chunks of text as large as newspaper articles and books. Thus, by application of Zipf's concept, complex constructions will be correspondingly rarer in appearance in a corpus. If this extension of Zipf's argument is valid and applicable here, as we claim, then the extension indicates a further dilution of complex syntactic structures in corpus, with a corresponding further skewing of percentage results.

Identifying and counting the number of different constructs begs the question in some ways, as there is needed an extremely large parsed corpus in which the rare constructs occur and are recognized. This would be an interesting project to perform in the Bank of English once the tagging (and preferably parsing) project is complete.

There is, however, another way we can approach this: by identifying the obvious easy constructs and simply counting those which are not handled.

A means by which the effects of repetitive or obvious constructs in corpora are removed from the scored percentages of parsers, appears appropriate, then. In order to do this, we wrote an extremely elementary (simple) grammar[3] (the rules for it are noted in Appendix A). We then counted the number of correct roles that this grammar gained, on a section of *Alice in Wonderland.* This simplistic parser scored a figure of 80.7% correct on the piece of text selected - which figure, in a sense, creates a baseline or zero-level for that piece of text. This is really the way that one traditionally develops a grammar - the initial version would normally be enhanced as further iterations are made, but whilst we were tempted to rectify the obvious problems, the statistics in Table 1 for this "obvious tagger" has not had the benefit of any refinement.

This "grammar" is designed to be used only for benchmarking, not as a real grammar or for a production parser. It can be employed in either of two ways: On the one hand, a comparison can be made between different corpora, by using the figures assessed by the grammar to characterize the difficulty of a corpus. Alternatively, and perhaps more interestingly, the figure calculated by the grammar on a corpus can be used to bring percentages of correct roles on words into a proportion which does have real relevance to a standard of parsing. For an example for this test, the statistics of say 97% of the total number of words correctly tagged by Tagger X on a corpus which has been assessed as having a zero-level of 80.7%, is re-balanced proportionately - in that case down to 84.45%: by which we mean that 84.45% of the words that are a test to a substantial extent, of a parsing program, are being handled correctly. It will be noted then that the grammar's rules leave plenty of scope to reward the parser which is operating even moderately well, and that the technique can be applied to both type I and type II errors.

The obvious parser is somewhat rough, but it makes the statistics offered by the measurement of correct word-roles on corpora far more meaningful *as long as the baseline is set for each corpus* - otherwise we are simply multiplying the error rate by 5 if we set the benchmark ceiling to (roughly) 20% rather than the more commonly assumed 100%.

Certainly, this re-balancing test means that a parser is no longer given credit for just correctly applying a tag which is completely obvious anyway (for example, the tag *article* for the word *the*), but that is currently what is occurring. And since *the* is usually the most common word in English text, that is usually occurring rather too often.

We conclude this section by noting that the baseline grammar's surprising success on the sentences

[3]Thought-time for creation of the "rules" for this "grammar" - about an hour, much of that hour typing and refining those rules. This "grammar" has nothing whatever to do with our own parser noted above.

for which the others failed, probably results both from rarer more complicated constructs not being recognized (which is the main point of it as a base-line) and from the fact that the grammar is the sole influence on the result (there is no probabilistic bias or arbitrary selection or omission of roles on open class words).

7.2 A measurement of the difficulty-level of a corpus

Rather than one attempting to locate a base-line for measurement and comparison, it should be possible to measure the comparative difficulty of corpora in terms of problematic constructs. For this purpose, we chose one phenomenon as representative: the number of syntactic usages or subcategorizations of verbs. To do this, we first selected, at random, a point of the text, and then we counted off the next eighty verbs. We then counted the number of times a verb out of those eighty verbs was presented to us in a different form: that is, either the word had not appeared as a verb at all before or, as a verb, it had been in some manner, used differently syntactically. Thus the same verb could be counted twice if it had a different subcategorisation or syntactic variation in each of two uses in the extract; our suggestion being that probably any syntactic phenomenon could be examined for variation, and verbs, being so central to a sentence, would probably indicate variety as successfully as any other syntactic feature could.

We believe that a variety is indicated in the results that we recorded, on eighty verbs: see table 2. However, we caution that the base line for the verbs needs to be determined from a comprehensive and representative corpus of English, and this has not yet been done.

On examination following that analysis, we decided that the scientific paper was unusually varied in its style at the particular point that we attempted assessment, which may account for that high rating. Our own opinion is that these figures corresponded well with the amount of variety in each piece of writing. In particular, extracts four and five were, in our opinion, of a particularly staid and dry, repetitive style.

From this brief study, we suggest that it may not be difficult to assess the syntactic complexity of pieces of writing, and we intend to try each piece of writing on the parsers themselves to see if a correlation can be observed between our estimate of the complexity of the work, the variety amongst the verb-forms presented, and the responses of the parsers themselves.

Appendices

A The simple grammar (for the *Obvious Parser*)

- The text is examined for all capitalised words that do not start sentences. These words become the proper nouns, and are all classified as nouns.
- All words with one, single, obvious role were merely given that role. Thus, the articles, the unequivocal conjunctions (e.g. *because* but not *so*), and most of the adverbs, were all just given their obvious role.
- Some words with an obvious class and a secondary, less likely class, were given their obvious class only. The equivocal modals (*will, might, can* etc.) were all given the role of *modal*, along with the unequivocal modals (*would, should* etc.)
- Demonstratives, equivocal possessive pronouns: were not pronouns - they were determiners only, (unless the word already had a class, otherwise given herein).
- The words *how* and *so* and *what* were conjunctions.
- The word *that* was always a relative pronoun.
- All other words which may be relative pronouns were relative pronouns.
- The word following an article was classified as a noun (unless it already had a class, otherwise given herein).
- The word following an adjective was classified as a noun (unless it already had a class, otherwise given herein).
- The word following a noun was classified as a verb (unless it already had a class, otherwise given herein).
- The word following a verb was classified as a noun (unless it already had a class, otherwise given herein).
- All *-ly*-inflected words were classified as adverbs.
- The words *once, twice, etc.,* and *either* and *here* and *just* and *there* were adverbs.
- *very* was always to the left of an adjective.
- A preposition (equivocal or not) always began a prepositional phrase and the prepositional phrases were only ever of the form

Book number	*Book genre*	*Number of different verb-forms*
1	Fiction book	64
2	Fiction book	55
3	Research Scientific paper	59
4	Scientific book	44
5	Textbook	48

Table 2: Count of different verb-forms for various extracts

preposition article/determiner object-of-prepositional-phrase or *preposition object-of-prepositional-phrase*.

- The word *too* was an intensifier.
- The word *thought* was always an ed-inflected word.
- The word *to* was always an infinitive, coupled with the word following it.
- All parts of the verbs *to be* and *to have* (other than the infinite itself, dealt with in the rule last above) were treated as auxiliaries.
- All *ing*-inflected words were classified as participles.
- All *ed*-inflected words were classified:
  - in all cases, all part-tense forms of the English "strong-verbs" being treated as *ed*-inflected words (with, of course, the further advantages to parsing given by the diffentiated forms of preterite and past-particple forms of "strong-verbs".)
  - if an auxiliary was to its left, then as a participle (so forming a verb, with the auxiliary).
  - if otherwise than the last, then as a verb.
- And anything not classified (as well as, of course, anything improperly classified) by the above rules, was wrong.

This grammar and its SCHEME LISP implementation is available from the first author on request.

References

Lewis Carroll. *Alice's Adventures in Wonderland; and Through the looking glass and what Alice found there.* Oxford University Press, London, 1975.

Jim Entwisle and Michael Groves. A method of parsing English based on sentence form. In *International Conference on New Methods in Language Processing*. Centre for Computational Linguistics, 1994.

Karlssen. Constraint grammar as a framework for parsing unrestricted text. In *Proceedings of the 13th International Conference of Computational Linguistics,*, volume 3, pages pp.168–173, 1990.

David Powers. Applications and explanations of zipf's law. In *International Conference on New Methods in Language Processing*, 1998.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language.* Longman, Great Britain, 1993.

Christer Samuelsson and Atro Voutilainen. Comparing a linguistic and a stochastic tagger. In *Proceedings of the 17th International Conference of Computational Linguistics*, 1997.

M. R. Speigel. *Theory and Problems of Statistics.* McGraw Hill, 1972.

Pasi Tapanainen and Timo Jarvinen. Syntactic analysis of natual language using linguistic rules and corpus-based patterns. In *Proceedings of COLING-94*, Kyoto, Japan, 1994.

Atro Voutilainen. A sytax-based part-of-speech analyser. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, 1995.

Atro Voutilainen and Mikko Silvonen. A short introduction to engcg. http://www.lingsoft.fi/doc/engcg/intro/components.html, 1996.

Atro Voutilainen and Pasi Tapanainen. Ambiguity resolution in a reductionistic parser. In *Proceedings of EACL '93*, 1993.

George K. Zipf. Human behaviour and the principle of least effort. *A.W.*, 1947.