"NATURAL LANGUAGE TEXTS ARE NOT NECESSARILY GRAMMATICAL AND UNAMBIGUOUS OR EVEN COMPLETE."

Lance A. Miller

Behavioral Sciences and Linguistics Group
IBM Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

The EPISTLE system is being developed in a research project for exploring the feasibility of a variety of intelligent applications for the processing of business and office text (1-3; the authors of 3 are the project workers). Although ultimately intended functions include text generation (e.g., 4), present efforts focus on text analysis: developing the capability to take in essentially unconstrained business text and to output grammar and style critiques, on a sentence by sentence basis.

Briefly, we use a large on-line dictionary and a bottom-up parser in connection with an Augmented Phrase Structure Grammar (5) to obtain an approximately correct structural description of the surface text (e.g., we posit no transformations or recovery of deleted material to infer underlying "deep" structures). In this process we always try to force a single parse output, even in the presence of true ambiguity. Grammatical critiques are provided by having very strong grammar restrictions in an initial processing of the sentence; should the application of grammar rules fail to lead to the identification of a complete, syntactically correct, sentence, we then process the material a second time, adding other rules which essentially relax certain constraints, such as subject-verb number agreement, thereby permitting us to recognize a wide variety of true grammatical errors. The stylistic critiques are based on measurements of the detailed hierarchical structure descriptions produced by the parser, letting us detect a variety of stylistic characteristics judged by "experts" to be undesirable: too great a distance between subject and verb, too much embedding, unbalanced subject/predicate size, excessive negation or quantification, etc.

The text corpus used for system construction and testing is a set of some 400 business letters, mostly written by individuals from within various organizations to individuals outside those organizations. These letters, which consist of approximately 2300 sentences, were selected from a larger collection (about 2000 letters) as being representative of the wide variety of styles, tones, subject matter, purposes, lengths, factual content, and organization-type found in the overall population of business letters. A corpus differing in so many of the above features is also heterogeneous with respect to syntactic structures -- and therefore with respect to the grammatical capabilities that must be incorporated for correct recognition. However, it was one thing to be prepared for structural diversity; it was quite another thing to be faced with the fact that our business letters are not some small to moderate subset of grammatical phenomena. Rather, they include all of the common and most of the arcane constructions one could find in, say, Warriner and Griffith (6). For example, the very first sentence we tackled was 29 words long and began "How nice it was to receive your letter complimenting our Manager, Bud Handy, on his courtesy ...": we ran into extraposition, inversion, infinitive nominalization, gerund phrase, and appositive all within the first 13 words! A primary consequence of this rich jumble of syntactic scree was the frequent annoyance of being stopped dead in our processing tracks as our grammar revealed itself to be yet once more incomplete.

But it was not only the incompleteness of the grammar (for correct sentences) that gave us trouble: many words were not recognized, sometimes sentences were incomplete, other times they were truly ungrammatical (via normal abnormalities of grammar or via what appeared to be a rather thoughtless -- or at least uninformed -- scattering of apostrophes and semicolons within the text) and often we were faced not with our desired single parse but with many. These then are the situations which cried out for techniques either to keep processing going or, at least, to keep it alive long enough for it to scratch out detailed informative guesses at structure on the parsing floor before expiring.

The techniques for hardiness and robustness which we have developed in the two years of implementation, and particularly recently, are mostly specific to the five trouble situations referred to above. For (i) unrecognized words (words not in our 125K entry on-line dictionary) we check first either for initial capitalization or for an internal hyphen, presuming a proper name -- noun -- part of speech for the former and either noun or adjective for the latter. As we improve our dictionary processing, to support efficient affix-stripping and stem storage, we now plan to hypothesize parts of speech based upon, in particular, the outer suffixes (e.g., "ly" pretty conclusively establishes multi-syllabic words as adverbs). This more "intelligent" processing at the part-of-speech level is particularly important for avoiding multiple false parses.

For the two situations of either (ii) an incomplete grammar failing to process a complete grammatical sentence, or (iii) an actual incomplete sentence (sentence fragment), we are no able to output a single "best" structural description when the grammar can do no more (Jensen and Heidorn, forthcoming). This partial structure is "best" in the sense that it provides the largest and most continuous coverage of the input text string, and it also adheres to certain orderings of parts of speech and non-terminal constituents. Our experience with such structures is that they are quite often correct, always better than a "CANNOT PARSE" outcome, and appear to be fairly usable for style critiquing. In the future we believe more can be done with sentence fragments by assuming, first, they are simply to be conjoined to some element of the previous sentence, or, second, they are an elaboration of an immediately preceding element; in either case the partial structure output should provide sufficient information to "hook" the fragments in correctly.

For (iv) truly ungrammatical sentences, as mentioned previously, we introduce a second pass with a number of grammatical restrictions relaxed; should any complete sentence structure result we can determine which relaxations were responsible and thereby actually identify the class of ungrammaticality. From the point of view of useful applications, this is much more of a desirable user-oriented function than an internal robust recovery procedure. Nonetheless, from the point of view of the style critiques at the sentence and paragraph levels, this procedure assures the best possible starting point, despite "noise" in the input text.

Finally, (v) the situation of multiple parses is dealt with by two techniques. The first is the deliberate attempt to construct the grammar rules such that no more than a single parse can squeeze through in most situations; the second is the development of a metric which computes a real number for each parse, based on its structural features, with the decision rule simply being to choose the parse with the smallest number (7). Our experience with this metric is that it usually leads to selection of the best all-around parse; such errors as are made would seem to require semantic -- and even pragmatic -- information to be weighed in the metric, a capability presently beyond our means.

REFERENCES

1. Miller, Lance A. "Project EPISTLE: A system for the automatic analysis of business correspondence." Proceedings of the First Annual National Conference on Artificial Intelligence, Stanford University, August, 1980, 280-282.

2. Miller, Lance A., George E. Heidorn, and Karen Jensen "Text-Critiquing with the EPISTLE System: An Author's Aid to Better Syntax." AFIPS Proceedings of the National Computer Conference, Chicago, May 4-7, 1981, 649-655.

3. Heidorn, George E., Karen Jensen, Lance A. Miller, Roy J. Byrd, and Martin S. Chodorow "The EPISTLE Text-Critiquing System." IBM Systems Journal, to appear Fall, 1982.

4. Jensen, Karen "Computer Generation of Topic Paragraphs: Structure and Style". Paper presented at the ACL Session of LSA Annual Meeting, New York City, December, 1981 (IBM Research Report, 1982).

5. Heidorn, George E. "Augmented Phrase Structure Grammars". In B. Nash-Webber and R. Schank (Eds.), Theoretical Issues in Natural Language Processing, Association for Computational Linguistics, 1975.

6. Warriner, J. E. and F. Griffith English Grammar and Composition. New York: Harcourt, Brace and World, Inc., 1963.

7. Heidorn, George E. "Experience with an easily computed metric for ranking alternative parses". Presentation at the Association for Computational Linguistics Meeting, Toronto, Canada, June 17, 1982.