Chapter 5

Parse Fitting and Prose Fixing

Karen Jensen, George Heidorn, Lance Miller, and Yael Ravin

Abstract

Processing syntactically ill-formed language is an important mission of a text-critiquing system. This chapter discusses how ill-formed input is treated by Epistle, the forerunner of Critique. Misspellings are highlighted by a standard spelling checker; syntactic errors are detected and corrections are suggested; and stylistic infelicities are called to the user’s attention. Central to the processing strategy is the technique of fitted parsing. When the rules of a conventional syntactic grammar are unable to produce a parse for an input string, this technique can be used to produce a reasonable approximate parse that can serve as input to the remaining stages of processing.

This chapter describes the fitting process and gives examples of ill-formed language situations where it is called into play. It also shows how a fitted parse allows the system to carry on its text-critiquing mission where conventional grammars would fail either because of input problems or because of limitations in the grammars themselves.

This chapter is a shortened version of Jensen et al. 1983.

K. Jensen et al. (eds.), Natural Language Processing: The PLNL Approach
© Kluwer Academic Publishers 1993
5.1 Introduction

The Epistle system addresses the problems of grammar and style checking of texts written in ordinary English (letters, reports, and manuals, as opposed to novels, plays, and poems). It is this goal that involves us so intimately with the processing of ill-formed language. Grammar checking deals with such errors as disagreement in number between subject and verb; style checking calls attention to such infelicities as sentences that are too wordy or too complex. A standard spelling checker is also included.

Our grammar is written in PLNLP (Heidorn 1972), an augmented phrase structure language which is currently implemented in LISP/370. At this time the Epistle grammar uses syntactic, but not semantic, information. Access to an online standard dictionary with over 100,000 entries (Webster's Seventh New Collegiate Dictionary), including part-of-speech and some other syntactic information (such as transitivity of verbs), makes the system's vocabulary essentially unlimited. We test and improve the grammar by regularly running it on a database of 2254 sentences from 411 actual business letters. Most of these sentences are rather complicated; the longest contains 63 words, and the average length is 19.2 words.

Since the subset of English represented in business documents is very large, we need a very comprehensive grammar and robust parser. At the time of this writing (1983), the rules of the grammar produce a single approximate parse for almost 70% percent of input text, and a small number of multiple parses for another 16%. The grammar can always be improved and its coverage extended, but that coverage will never reach 100%. For those strings that cannot be fully parsed by rules of the grammar we use a heuristic best fit procedure that produces a reasonable parse structure.

5.2 The fitting procedure

The fitting procedure begins after the grammar rules have been applied in a bottom-up, parallel fashion, but have failed to produce an S node that covers the string. At this point, as a by-product of bottom-up parsing, records are available for inspection that describe the various segments of the input string from many perspectives, according to the rules that have been applied. The term fitting has to do with selecting and fitting these pieces of the analysis together in a reasonable fashion.

The fitting algorithm, which is itself implemented as a set of PLNLP rules, proceeds in two main stages: first, a head constituent is chosen; next, remaining