Chapter 8

ROBUST PARSING AND BEYOND

Jean-Pierre Chanod
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, France
Jean-Pierre.Chanod@xrce.xerox.com
http://www.xrce.xerox.com/

Abstract  Robust analysis covers not only processing partially ungrammatical input (Carbonell & Hayes 1983, Weng 1993), but more generally unrestricted text as actually produced by end-users in various situations.

While early work in robust analysis dates back to the seventies, the topic has recently gained broader attention, due to the large quantities of textual data now available in electronic form (e.g. technical documentation, news, web pages, email, voice system).

This article first introduces the notion of robust parsing in natural language processing. It then reviews a series of linguistic phenomena that illustrate the need for specific approaches to robust parsing. Finally, it mentions some of pioneer and recent work in robust parsing.

8.1 Introduction

The growing need to process large quantities of text and speech now widely available in electronic format has boosted the effort to produce robust language processing systems.

In that context, corpus linguistics has gained deeper attention in the language processing community (Garside, Leech & Sampson 1987, Wilks 1993). The most spectacular effect of this shift has been the push of statistical-based methods, with speech recognition, part-of-speech taggers, information retrieval systems reaching higher and higher levels of accuracy (Church 1988, Brown, Della-Pietra, Jelinek, Mercer & Roossin 1988, Briscoe & N. Waegner 1992, Cutting, Kupiec, Pedersen & Sibun 1992, Charniak 1993).

In contrast, such characteristics underline the weaknesses of more traditional linguistic approaches, especially their inability to process large quantities
of data and, as a natural consequence, their lack of consideration for evaluation measures.

Still robustness is not about statistical vs. rule-based methods. It is not about virtual core languages, be they defined by principles or by frequency counts. It is not even about the quantity of unrestricted text that can be parsed by a given system. Robustness is about exploring all constructions humans actually produce, be they grammatical, conformant to formal models, frequent or not. Linguistic phenomena, regardless of their oddity or frequency, account for the meaning of whatever segment of text they appear in. Rare or deviant phenomena may become predominant in certain domains or genres (e.g. telegraphic style in maintenance reports or news dispatches, dates and addresses in agenda management applications, imperative elliptic statements in user interfaces, lists of complex noun phrases in medical prescriptions, etc.).

In this view, robustness is a matter of breadth and depth of linguistic analysis. Altogether.

### 8.2 Linguistic Descriptions and Robustness

This section addresses the issue of linguistic descriptions confronted with robustness. It investigates how the variety and complexity of linguistic constructions found in unrestricted texts may reduce the domain of validity of NLP descriptions.

In order to illustrate this purpose, a few characteristic examples are selected and discussed, where linguistically motivated propositions are confronted with a series of surface phenomena. Linguistic phenomena in situ (i.e. in corpora) undergo multiple variations and distortions that core descriptions do not capture easily. This does not necessarily undermine the value of principles from a pure linguistic perspective (Chomsky 1981, Berwick 1991), but it shows that linguistically sound principles do not necessarily translate into effective computational models (Jensen 1988, Nagao 1992). Such primary facts, on the other hand, are essential in the conception of robust systems.

#### 8.2.1 Robustness and Unification of Lexical Features

Unification of lexical features has been a productive research paradigm in syntax and has led to various formalisms such as GPSG, LFG or TAG (Joshi, Levy & Takahashi 1975, Kaplan & Bresnan 1982, Gazdar, Klein, Pullum & Sag 1985). Unification presents a computationally elegant way of handling such phenomena as agreement or subcategorisation. Nevertheless, linguistic facts collected across languages present a wide variety of situations where the underlying principles behind unification do not properly account for agreement or verb subcategorisation. This does not reduce the beauty of the underlying principles, but creates an obstacle to the implementation of robust parsers. Indeed, one basic requirement for a robust parser is to handle standard phenomena