# Example-Based Parsing for Resource-Deficient Languages

By

Shailly Goyal (Nee Kansal)

Department of Mathematics

*Submitted in fulfillment of the requirements of the degree of Doctor of Philosophy*

*to the*


Indian Institute of Technology Delhi
June 2007

I. I. T. DELHI.
LIBRARY
Acc. No. TH-3506

I.I.T. DELHI
PROCURED
LIBRARY

TH
81322
GOV-E

*To my dear husband . . .*

*a token for your love and inspiration*

# Certificate

This is to certify that the thesis entitled *Example-Based Parsing for Resource-Deficient Languages* submitted by *Ms. Shailly Goyal (Nee Kansal)* to the Indian Institute of Technology Delhi, for the award of the Degree of Doctor of Philosophy, is a record of the original bona fide research work carried out by her under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

New Delhi
June 2007

**Dr. Niladri Chatterjee**
**Supervisor**

# Acknowledgements

This thesis would not have been possible without support of many people. It is impossible to acknowledge all of them here, but there are few people who deserve a special mention.

I am forever indebted to my Ph.D. supervisor, Dr. Niladri Chatterjee, for teaching me the basics of research. I am grateful for his unfailing support throughout this work. He was great in letting me do what I wanted, gave consistently good advice, and encouraged me when the things were not going fine. He was extremely patient with my failings, and his trust in my abilities helped me overcome many difficulties.

I thank IIT Delhi for providing me teaching assistantship and the necessary facilities required to pursue my research smoothly. I am thankful to Prof. B. Chandra, Head of the Deaprtment and one of my SRC member, for her help. Special thanks are due to Prof. S. Kaushik, the other member of my SRC, for the constructive suggestions provided at various stages of this research. I am grateful to Dr. N. Nataraj for motivating me to join the Ph.D. Many thanks to my pre-Ph.D. course instructors - Prof. S. R. K. Iyenger, Prof. J. B. Srivastava and Dr. A. Tripathi - for helping me strengthen my basics. I express my regards to Prof. R. K. Sharma, DRC Chairperson, for his support. Thanks to non-teaching staff for their assistance.

I am very lucky to have many good friends and colleagues who made my experience in this department greatly enjoyable. Thanks to *Sary* and *Annie* for being

there to share the various highs and lows I faced during this research. I thank them also for the many fun moments I spent with them. I appreciate my seniors, Deepa, Sonia and Prabhakar, for making me comfortable in the department, and motivating in the initial days of my research. I will always remember Manju, Pratibha, Kanchan, Preeti, Pandey*ji*, Megha, Reshma, Geeta, Vaneeta, *Dinu* and all others for providing fun-filled and cheerful environment in the department. Work looked less mundane with all of them around.

I will never forget Krishna for his genuine concern. He always boosted and supported me selflessly during the whole process. I thank him also for countless impromptu academic and non-academic discussions which we had, from which I have learnt a lot about research, in particular, and life, in general. Just saying thanks will be too little a gratitude towards him.

Finally, and most importantly, thanks to many wonderful people in my personal life for giving their unwavering support throughout, especially during this work. No words can be ever enough to thank my husband, Nilesh, for his constant encouragement, understanding and endless patience. My Ph.D. had always been first preference for him, and he kept everything else on hold for this. I am grateful to my in-laws, Prof. R. K. Goyal and Mrs. Usha Goyal, for their whole-hearted support, and for providing very encouraging environment at home to pursue my research. I convey my earnest thanks to my sister-in-laws, Dr. Reena Mittal and Dr. Neena Jain, for their moral support. I would like to express my deepest gratitude towards *nanaji*, Mr. O. P. Mathuria, and *dadaji*, Mr. S. L. Goyal, for their blessings. Sincere thanks are due to my parents, Er. R. K. Kansal and Mrs. Manju Kansal, for their unending love and blessings. My siblings, Nidhi and Utkarsh, deserve a special mention for cheering me always.

Above all, I thank God for making this thesis possible.

New Delhi

Shailly Goyal (Nee Kansal)

# Abstract

Aim of the present research is to develop parsing schemes for natural language sentences. Parsed corpus is essential for various natural language processing (NLP) activities, but its availability cannot be guaranteed for most of the languages. Furthermore, development of parsed corpus or parser is not an easy task using traditional approaches, viz. rule-based and statistical. This is because success of these approaches almost invariably demands a huge amount of computational resources that are not typically available for most of the natural languages. We feel that example-based (EB) approaches can serve as suitable alternatives at this juncture. The major advantage of these approaches is that their demand on computational resources is much less in comparison with the traditional approaches, yet EB approaches are useful in developing robust techniques as is envisaged in many areas of artificial intelligence, NLP in particular. In this work we have pursued following two aspects of example-based parsing:

*Bilingual Parsing.* In this methodology a sentence is parsed using the parse of its parallel sentence. While projecting the syntactic relations from one langauge to another, we have considered similarities as well as dissimilarities between the two languages. Hence we are able to develop generalized schemes that can work on a wide variety of source-target language pairs. We have developed parsing schemes for simple as well as complex sentences.

*Monolingual Parsing.* In this scheme a sentence is parsed using the parse knowledge of examples of the same language. We have developed schemes for parsing sentences of a language by acquiring appropriate knowledge from a parsed example base of the same language. In this work we have devised ways to take care of various problems, such as unknown words, free word order property, morphological variations, effectively.

We have developed both these schemes in a generalized way with minimal dependence on linguistic knowledge so that the schemes developed can be used across a wide spectrum of languages. In this work we have done a thorough case-study on Hindi. For nitty-gritty details of the parsing schemes, where linguistic details are inevitable, we have considered English and Hindi as the source and target language, respectively. In this work we have chosen link grammar as the underlying grammar for representing the parse of sentences. One fundamental requirement therefore is a link grammar for languages under consideration. Since, no such grammar exists for Hindi, we have developed a link grammar for Hindi. For this task also we follow example-based approach.

*Development of Hindi Link Grammar.* Instead of developing the link grammar for Hindi from scratch, in this work we have made appropriate modifications in the English link grammar to suit the requirements of the Hindi grammar. We have shown how English links can be adapted for Hindi by taking care of its various grammatical nuances (e.g. free word order, noun and verb morphology, influence of subject/object on verb morphology) that make Hindi grammar distinctly different from English.

The parsing schemes developed in this work have been implemented, and tested on a reasonably sized example base. Still we have been able to demonstrate clearly the efficacy of these schemes. We feel that our research will pave the way for quick development of parsers for other languages.

# Contents