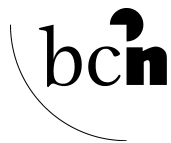


# Using ILP to Learn Local Linguistic Structures

Stasinos Themistokleous Konstantopoulos



The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen.



Groningen Dissertations in Linguistics 45  
ISSN 0928-0030

Document prepared with  $\text{\LaTeX} 2_{\epsilon}$  and typeset by pdf $\text{\TeX}$ ,  
Greek summary prepared with Lambda and typeset by Omega.  
© 2003 Stasinou Konstantopoulos

Rijks*universiteit* Groningen

## Using ILP to Learn Local Linguistic Structures

Proefschrift

ter verkrijging van het doctoraat in de  
Wiskunde en Natuurwetenschappen  
aan de Rijksuniversiteit Groningen  
op gezag van de  
Rector Magnificus, dr. F. Zwarts,  
in het openbaar te verdedigen op  
vrijdag 28 november 2003  
om 14.15 uur

door

Stasinos Themistokleous Konstantopoulos

geboren op 2 september 1973  
te Athene, Griekenland

*iv*

Promotor: Prof. dr. ir. J. Nerbonne

Beoordelingscommissie: Prof. dr. W. Daelemans  
Prof. dr. T. Kuipers  
Prof. dr. G. R. Renardel de Lavalette

*We in the West have no syllogism exactly equal to the anumana and it is a shame that we do not, because had we such a rigorous form by which to check our inductive reasoning, Bishop Timothy Archer might well know of it, and had he known of it he would have known that his mistress waking up to find her hair singed does not, in fact, prove that the spirit of his dead son has returned from the other world.*

*Philip K. Dick, The Transmigration of Timothy Archer*



# Preface

This thesis would not have been possible without the help of many people, to all of whom I'm deeply indebted. First of all I would like to thank John Nerbonne for supervising my project. Without his experience and guidance it would have taken a tougher and longer time, and the result would have been much less satisfactory. And particularly because without his tireless instigations to clearly and fully explain my thoughts on paper, this thesis would have been impossible to read.

Then I would like to thank all the people in Alfa-Informatica in particular and the H.1311 corridor in general for all the ideas we have exchanged and all the coffees we've had together. (I hope I got the corridor number right, one can never be sure with the Harmonie building's numbering scheme, but anyway you know which corridor I mean.) I would like to single out Tony Mullen and Rob Malouf, for all the stimulating discussions we have had over coffee or beer. And for being very good office-mates; although on that particular front I have to say that I have nothing but good memories from my current office-mates as well, Begoña Villada Moirón and Susanne Schoof.

I would also like to thankfully mention Ashwin Srinivasan for writing Aleph. And all the people outside the department who have shown interest in my project and with whom I have had fruitful discussions and especially Rui Camacho, Vitor Santos and all the people in Porto. I would also like to thank Pavel Brazdil, the director of the AI lab in Porto, for offering me hospitality at the lab.

On the non-computational side of my project, lots of thanks to Wouter Jansen, Dicky Gilbers, and Roberto Bologniesi for every discussion we have had about Phonology and every bit of ignorance and lack of linguistic background they have helped dissolve. And to the Zeppelin for helping me clarify some Logic Programming concepts in my head last March. And to Kengo Harimoto for his explanations and bibliographical references to Indian Logic.

It would also like to thank the members of my reading committee, Walter Daelemans, Theo Kuipers, and Gerard Renardel de Lavalette; their comments and suggestions have had a great impact at improving the final version

of this thesis.

My gratitude also goes to Rob Visser, Anna Hausdorf, Wyke van der Meer and the *Secretariaat CNL* for their administrative support. And Shoji Yoshikawa in *Letteren* and Kees Visser in HPC for their systems administration: their contribution to creating a smooth working environment is very much appreciated.

There is also lots of people whom I am indebted to, although they were not directly related to my project or to this thesis. First of all I should thank Willem Moolenburgh for my being here in the first place: his enthusiastic description of life in Groningen played a great role in overcoming my original reservations about making such a long-term commitment to a small town. And then, of course, all the people who made Groningen live up to Willem's enthusiasm starting from Tony and Ivo who were the first people I met here; all the people who have joined me for a cup of coffee and a smoke at the last table to the right, on the sunny side of the smoking section of the Harmonie Kantine; my flatmates from 4Zuid to Kl. Raamstraat to Oosterparkwijk; Алёна, Kanat, Laura, Марина, Simo, the Wrocław students of Dutch, the ex-YUs, two generations of EMCL and Euro-Culture students, and all the people I've met in Groningen. And, of course, a special mention should go to *Vera*, *Filmcentrum Poelenstraat*, and the *Paard van Troje*: Groningen just simply wouldn't have been the same without these fine establishments.

A special thanks goes to Leonoor van der Beek and Eleonora Rossi, my paranimfs, for their help with organizing everything so nicely. Well, strictly speaking for their agreeing to organize everything so nicely in the near future, but I'm sure they'll do a great job. Thanks once again to Leonoor for by far the best *samenvatting* translation I've come across in my 5 years here, but also for her comments that greatly improved the English original as well.

And, of course, a big thanks to Θεμιστοκλή and Ελευθερία, my parents, for their support, and for my being here in the first place!

Finally, I would like to close with a somewhat unrelated comment: if there is one single thing the whole Groningen experience taught me, it's that the Ukrainian redberry is the sweetest and juiciest fruit there is.

That was it. If I forgot somebody, my most profound and sincere apologies, I didn't mean to!

I hope you'll enjoy reading my thesis.

Groningen, 11-10-2003

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview . . . . .	2
1.2	Deduction and Induction . . . . .	3
1.2.1	Aristotelian Logic . . . . .	4
1.2.2	Rationalism and Empiricism . . . . .	5
1.3	Inductive Inference Operators . . . . .	6
1.4	Induction and Justification . . . . .	8
1.5	Perfect Induction . . . . .	9
<b>2</b>	<b>Inductive Logic Programming</b>	<b>11</b>
2.1	Decision Tree Induction . . . . .	11
2.1.1	The Original Concept-Learning Systems . . . . .	14
2.1.2	Entropy as Search Heuristic . . . . .	16
2.1.3	Avoiding Overfitting . . . . .	17
2.1.4	Some Thoughts... . . . .	18
2.2	Propositional Rule Induction . . . . .	18
2.2.1	Learning Lists of Rules . . . . .	21
2.2.2	Traversal Operators . . . . .	22
2.2.3	Evaluating Rule Performance . . . . .	23
2.2.4	Example-Seeded Search . . . . .	27
2.2.5	Theory-Level Post-Processing . . . . .	27
2.2.6	Some More Thoughts... . . . .	28
2.3	Introducing ILP . . . . .	28
2.3.1	The Task of ILP . . . . .	29
2.3.1.1	Formalising in Normal Semantics . . . . .	30
2.3.1.2	Formalising in Definite Semantics . . . . .	33
2.3.2	Inverse Entailment . . . . .	34
2.3.2.1	Specialization and Generalization Operators . . . . .	34
2.3.2.2	$\theta$ -subsumption . . . . .	35
2.3.2.3	Resolution . . . . .	37
2.3.3	Prior Knowledge . . . . .	38

2.3.4	The Evaluation Function . . . . .	41
2.4	The Progol Algorithm . . . . .	43
2.4.1	Saturation . . . . .	43
2.4.2	Reduction . . . . .	46
2.4.3	Cover removal . . . . .	50
2.5	Other Approaches to ILP . . . . .	50
2.5.1	Theory-Level Search . . . . .	50
2.5.2	Background Knowledge Revision . . . . .	51
2.6	Why ILP? . . . . .	52
<b>3</b>	<b>Data-Parallel ILP</b>	<b>53</b>
3.1	The Message Passing Interface . . . . .	54
3.1.1	Basic MPI Concepts . . . . .	54
3.1.2	Some More MPI Functions . . . . .	57
3.2	The Yap/MPI Interface . . . . .	59
3.2.1	Prolog Term Messages . . . . .	60
3.2.2	Point-to-Point Communication . . . . .	60
3.2.3	Broadcasting . . . . .	63
3.3	Evaluating Clauses in Parallel . . . . .	64
3.3.1	Loading the Examples . . . . .	64
3.3.2	Proving the Examples . . . . .	65
3.4	Testing Aleph/MPI . . . . .	68
3.4.1	Learning the odd numbers . . . . .	69
3.4.2	Other Approaches . . . . .	71
3.4.3	Conclusions . . . . .	71
<b>4</b>	<b>Shallow Parsing</b>	<b>73</b>
4.1	Full vs. Shallow Parsing . . . . .	73
4.2	Chunking . . . . .	76
4.2.1	What is a chunk? . . . . .	76
4.2.2	Noun Phrase Chunks . . . . .	78
4.3	Chunking as Tagging . . . . .	78
4.3.1	Bracket Tagging . . . . .	79
4.3.2	Inside/Outside Tagging . . . . .	79
4.3.3	Comparison . . . . .	80
4.4	Inducing a BaseNP Chunker . . . . .	81
4.4.1	Experimental Setup . . . . .	82
4.4.2	The Dataset . . . . .	83
4.4.3	List-Access Background Predicates . . . . .	84
4.4.4	List-Manipulation Background Predicates . . . . .	86
4.4.5	Linguistic Background . . . . .	86

4.4.6	The Baseline Theory . . . . .	87
4.4.7	Prior Bias . . . . .	88
4.5	Results and Conclusions . . . . .	88
4.5.1	Cascades of Chunkers . . . . .	90
<b>5</b>	<b>Phonotactics</b>	<b>93</b>
5.1	Extracting the Data . . . . .	93
5.2	The Background Knowledge . . . . .	96
5.3	The Baseline theory . . . . .	97
5.4	The IPA Chart . . . . .	97
5.4.1	Design . . . . .	98
5.4.2	Results . . . . .	99
5.5	Feature Classes . . . . .	100
5.5.1	Design . . . . .	100
5.5.2	Results . . . . .	101
5.6	Sonority Scale . . . . .	104
5.6.1	Design . . . . .	104
5.6.2	Implementation and Results . . . . .	105
5.7	The Search Space . . . . .	106
5.7.1	Bottom Clause Size . . . . .	106
5.7.2	Search-Space Size . . . . .	108
5.7.3	Data-Parallelism Vs. Or-Parallelism . . . . .	109
5.8	Conclusions . . . . .	110
<b>6</b>	<b>Conclusion</b>	<b>113</b>
6.1	Computational Complexity . . . . .	113
6.2	Summation of Results . . . . .	114
6.3	Discussion . . . . .	116
6.4	Future Directions . . . . .	118
	<b>Summary in Dutch</b>	<b>121</b>
	<b>Summary in Greek</b>	<b>125</b>
	<b>Bibliography</b>	<b>129</b>
	<b>Groningen Dissertations in Linguistics</b>	<b>139</b>

