

# Machine Learning of Phonotactics

Erik F. Tjong Kim Sang



The research described in this thesis has been made possible by a grant from the Dutch Research School in Logic (formerly *Nederlands Netwerk voor Taal, Logica en Informatie*).

GRONINGEN DISSERTATIONS IN LINGUISTICS 26

ISSN 0928-0030

RIJKSUNIVERSITEIT GRONINGEN

# Machine Learning of Phonotactics

Proefschrift

ter verkrijging van het doctoraat in de  
Letteren  
aan de Rijksuniversiteit Groningen  
op gezag van de  
Rector Magnificus dr. D.F.J. Bosscher,  
in het openbaar te verdedigen op  
maandag 19 oktober 1998  
om 16.15 uur

door

Erik Fajoen Tjong-Kim-Sang

geboren op 6 december 1966  
te Utrecht

Promotor: Prof. Dr. Ir. J.A. Nerbonne

# Acknowledgements

It is customary in my research group to add to one's PhD thesis a section with reflections about life during the production of this book and acknowledgements to people who have been friendly to the author during this period. I will make no exception to this custom. Over the past eight years and twenty one days I have, with some interruptions, worked on a research project about the application of machine learning methods in natural language processing. Over these years I have learned a lot but I have also found out that there is a lot left to be learned. My average writing speed of nineteen words per day has resulted in the PhD thesis on this subject which I am submitting right now.

I want to thank Frans Zwarts and Jan de Vuyst for starting this research project and giving me the opportunity to work in it. I am also grateful to the organization Nederlands Netwerk voor Taal, Logica en Informatie (currently Dutch Research School in Logic), which has supplied the grant that has made my research project possible. John Nerbonne has played a major role in the fact that this the project has resulted in a PhD thesis. He has been responsible for the project supervision including the initial suggestion of the final thesis topic. John also created the possibility for me to stay longer at the department of Alfa-informatica than the time offered by my original project. Thanks!

My colleagues of the department Alfa-informatica of the University of Groningen in The Netherlands were responsible for creating a pleasant working environment during the five years I was allowed to spend with them. For that reason I want to thank Bert Bos, Dicky Gilbers, Erik Kleyn, Garry Wiersema, George Welling, Gertjan van Noord, Gosse Bouma, Harry Gaylord, Joop Houtman, Mark-Jan Nederhof, Mettina Veenstra, Peter Blok, Petra Smit, Rob Koeling, Shoji Yoshikawa and Yvonne Vogelzang. I am specially indebted to Mettina and Bert. I was fortunate enough to spend most of my Groningen years working in the same room as them. They played a major role in my development during these years, both on a scientific and a personal level. Thanks!

After Groningen I have spent three years at the department of Linguistics of Uppsala University in Sweden. I want to thank Anna Sångvall Hein for offering me the opportunity to work in what I have regarded as a demanding but instructive environment. During my years in Uppsala I have been fortunate enough to work with the following people: Annelie Borg-Bishop, Bengt Dahlqvist, Gunilla Fredriksson, Hong Liang Qiao, Jon Brewer, Jörg Tiedemann, Klas Prytz, Lars Borin, Leif-Jöran Olsson, Malgorzata Stys, Mariana Damova, Mark Lee, Mats Dahllöf, Olga Wedbjer Rambell, Per Starbäck and Torbjörn Lager. I want to reserve a special word of gratitude to my former students in Uppsala which as a group have been a great example of motivation towards their studies and general friendliness both inside and outside the classroom. Thanks!

No thesis can be finalized without being read by and approved by a thesis committee. I want to thank Anton Nijholt, Ger de Haan and Nicolay Petkov for being part of my thesis committee and for reading and commenting the thesis.

In the past eight years I have met many nice people in different circumstances both in Groningen, Uppsala and at other locations. Their kindness has had a positive influence on my general well-behavior which has contributed to the completion of this thesis. I want to thank all of them.

Groningen, August 21, 1998,

Erik Tjong Kim Sang

# Contents

<b>Acknowledgements</b>	<b>5</b>
<b>Contents</b>	<b>7</b>
<b>1 Introduction</b>	<b>11</b>
1 Theoretical background . . . . .	12
1.1 Problem description . . . . .	12
1.2 Data representation . . . . .	13
1.3 Positive and negative learning examples . . . . .	14
1.4 Innate knowledge . . . . .	14
2 Experiment setup . . . . .	15
2.1 Goals . . . . .	16
2.2 The training and test data . . . . .	17
2.3 Data complexity . . . . .	18
2.4 The linguistic initialization model . . . . .	20
2.5 Elementary statistics . . . . .	22
3 Related work . . . . .	23
3.1 The work by Ellison . . . . .	23
3.2 The work by Daelemans et al. . . . .	24
3.3 Other work . . . . .	25
<b>2 Statistical Learning</b>	<b>27</b>
1 Markov models . . . . .	27
1.1 General description of Markov models . . . . .	27
1.2 The forward procedure . . . . .	29
1.3 The Viterbi algorithm . . . . .	31
2 Hidden Markov Models . . . . .	34
2.1 General description of Hidden Markov Models . . . . .	34
2.2 The extended forward procedure . . . . .	35
2.3 The extended Viterbi algorithm . . . . .	37
2.4 Learning in a Hidden Markov Model . . . . .	39
2.5 Using Hidden Markov Models in practice . . . . .	44

3	Initial Experiments . . . . .	45
3.1	A test experiment . . . . .	45
3.2	Orthographic data with random initialization . . . . .	46
3.3	Orthographic data with linguistic initialization . . . . .	48
3.4	Discussion . . . . .	50
4	Experiments with bigram HMMs . . . . .	52
4.1	General bigram HMM experiment set-up . . . . .	53
4.2	Orthographic data with random initialization . . . . .	53
4.3	Orthographic data with linguistic initialization . . . . .	55
4.4	Phonetic data with random initialization . . . . .	58
4.5	Phonetic data with linguistic initialization . . . . .	60
5	Concluding remarks . . . . .	63
<b>3</b>	<b>Connectionist Learning</b>	<b>65</b>
1	Feed-forward networks . . . . .	65
1.1	General description of feed-forward networks . . . . .	66
1.2	Learning in a feed-forward network . . . . .	68
1.3	Representing non-numeric data in a neural network . . . . .	71
2	The Simple Recurrent Network (SRN) . . . . .	72
2.1	General description of SRNs . . . . .	73
2.2	Learning in SRNs . . . . .	75
2.3	Using SRNs for language experiments . . . . .	76
3	Experiments with SRNs . . . . .	77
3.1	General experiment set-up . . . . .	78
3.2	Finding network parameters with restricted data . . . . .	80
3.3	Orthographic data with random initialization . . . . .	83
3.4	Orthographic data with linguistic initialization . . . . .	85
4	Discovering the problem . . . . .	87
4.1	The influence of the number of valid successors of a string . . . . .	87
4.2	Can we scale up the Cleeremans et al. experiment? . . . . .	88
4.3	A possible solution: IT-SRNs . . . . .	90
4.4	Experiments with IT-SRNs . . . . .	91
5	Concluding remarks . . . . .	93
<b>4</b>	<b>Rule-based Learning</b>	<b>95</b>
1	Introduction to Rule-based Learning . . . . .	95
1.1	Positive versus negative examples . . . . .	96
1.2	The expected output of the learning method . . . . .	97
1.3	Available symbolic learning methods . . . . .	97
2	Inductive Logic Programming . . . . .	99
2.1	Introduction to Inductive Logic Programming . . . . .	99
2.2	The background knowledge and the hypotheses . . . . .	102
2.3	Deriving hypotheses . . . . .	105
2.4	The hypothesis models and grammar theory . . . . .	107

3	Experiments with Inductive Logic Programming . . . . .	110
3.1	General experiment setup . . . . .	110
3.2	Handling orthographic and phonetic data . . . . .	112
3.3	Adding extra linguistic constraints . . . . .	113
3.4	Discussion . . . . .	116
4	Alternative rule-based models . . . . .	118
4.1	Extending the model . . . . .	118
4.2	Deriving extended hypotheses . . . . .	120
4.3	Experiments with the extended model . . . . .	122
4.4	Compressing the models . . . . .	124
5	Concluding Remarks . . . . .	128
<b>5</b>	<b>Concluding remarks</b>	<b>131</b>
1	Experiment results . . . . .	131
2	Recent related work . . . . .	134
3	Future work . . . . .	136
	<b>Bibliography</b>	<b>139</b>
	<b>Samenvatting</b>	<b>145</b>