

João Luís Garcia Rosa

**UM SISTEMA HÍBRIDO SIMBÓLICO-CONEXIONISTA PARA O
PROCESSAMENTO DE PAPÉIS TEMÁTICOS**

**Tese apresentada ao Curso de Lingüística do Instituto de
Estudos da Linguagem da Universidade Estadual de
Campinas como requisito parcial para obtenção do título
de Doutor em Lingüística**

Orientador: Prof. Dr. Edson Françaço

Co-orientador: Prof. Dr. Márcio Luiz de Andrade Netto

Campinas
Instituto de Estudos da Linguagem
1999

Prof. Dr. Edson Françaço – Orientador

Prof. Dr. Márcio Luiz de Andrade Netto – Co-orientador

Profa. Dra. Márcia Maria Cançado Lima

Profa. Dra. Eleonora Cavalcante Albano

Prof. Dr. Fernando Antonio Campos Gomide

Prof. Dr. Plínio Almeida Barbosa

Campinas, Junho de 1999

Dedicatória

Esta Tese é dedicada a

Mariana e Susy, Futuro e Presente.

Agradecimentos

Com o perdão de omissões tão certas quanto indesejáveis, agradeço às seguintes pessoas, sem as quais certamente não teria alcançado este meu objetivo:

- Meus pais, que mostraram o caminho correto a seguir nesta vida e sempre estiveram presentes;
- Minha esposa, cujo companheirismo e compreensão viabilizaram esta minha longa jornada;
- Minha filha, apesar de ainda não compreender a importância deste trabalho, pela sua pouca idade, adotou a postura de apoio e carinho, própria de um adulto;
- Os professores do Instituto de Estudos da Linguagem da Unicamp, em especial, Eleonora Albano, Plínio Almeida Barbosa, Ester Míriam Scarpa, Mary Kato e Rodolfo Ilari, pelas sugestões valiosas e pelo encaminhamento adequado dos meus trabalhos de Qualificação;
- A professora da Universidade Federal de Minas Gerais, Márcia Cançado, que apesar de distante geograficamente, esteve próxima através de seus exemplos e orientação;
- Os professores do Instituto de Informática da Pontifícia Universidade Católica de Campinas, em especial sua ex-diretora, Angela de Mendonça Engelbrecht, cujo apoio durante todo o desenvolvimento da Tese, possibilitou a sua finalização;
- Os alunos e colegas do Instituto de Estudos da Linguagem da Unicamp e do Instituto de Informática da PUC-Campinas, pelo apoio anônimo e incessante;
- O professor da Faculdade de Engenharia Elétrica e de Computação da Unicamp, Márcio Luiz de Andrade Netto, pela co-orientação atuante durante todo o meu trabalho;
- O professor Edson Françoze, meu orientador, que sempre confiante, ajudou-me a superar as muitas barreiras, naturais de uma área nova e desafiante para um ex-aluno de engenharia, e assim conquistar este título multidisciplinar.

Epígrafe

“O processamento de informação adicional deve ser capaz de manipular o japonês, o inglês e outras línguas naturais. Este é um dos temas centrais da Inteligência Artificial e ao mesmo tempo, uma área da Lingüística: relações intensas existem entre a Lingüística e os computadores.”

Kazuhiro Fuchi, *Fifth Generation Computers: Some Theoretical Issues*, 1984.

Sumário:

Capítulos

CAPÍTULO 1 - INTRODUÇÃO.....12

- 1.1 Considerações Iniciais
- 1.2 A Abordagem Simbólica
- 1.3 A Abordagem Conexionista
- 1.4 A Abordagem Híbrida Simbólico-conexionista

CAPÍTULO 2 - PROCESSAMENTO DE LINGUAGEM NATURAL: DA ENGENHARIA DE COMPUTAÇÃO À LINGÜÍSTICA.....16

- 2.1 Introdução
- 2.2 O Sistema P3A
 - 2.2.1 A Análise da Forma
 - 2.2.2 A Análise do Significado
 - 2.2.3 A Análise Temporal
 - 2.2.4 Arquitetura do Modelo
 - 2.2.5 Microcaracterísticas Semânticas
 - 2.2.6 Unidades de Estrutura de Sentença
 - 2.2.7 Representação de Papel de Caso
 - 2.2.8 Detalhes do Processamento de Sentença e Aprendizado
 - 2.2.9 Experimentos de Simulação
 - 2.2.10 Conclusão
- 2.3 O Sistema CPPro
 - 2.3.1 As Metas do CPPro
 - 2.3.2 Experimentos de Simulação
 - 2.3.3 Conclusão
- 2.4 Considerações sobre os Sistemas
- 2.5 Conclusão Geral

CAPÍTULO 3 - REPRESENTAÇÕES DISTRIBUÍDAS NOS SISTEMAS HÍBRIDOS PARA O PLN.....40

- 3.1 Os Sistemas Conexionistas que Fazem PLN
- 3.2 Microcaracterísticas Semânticas
 - 3.2.1 Contexto
 - 3.2.2 Microcaracterísticas e Contexto

3.2.3 Mecanismos do Processamento da Sentença	
3.3 Representações Distribuídas	
3.4 A Abordagem Híbrida	
3.5 Críticas ao Conexionismo	

CAPÍTULO 4 - ASPECTOS ESTRUTURAIS E SEMÂNTICOS DOS PAPÉIS TEMÁTICOS 55

4.1 Introdução aos Papéis Temáticos	
4.2 O Papel Temático na Teoria Gerativa	
4.2.1 O Papel Temático na GB	
4.2.2 Um Modelo Computacional Baseado em Princípios	
4.3 A Visão Semântica dos Papéis Temáticos	
4.3.1 A Interface Sintaxe-Semântica	
4.3.2 O Problema Lingüístico dos Ergativos	
4.4 A Natureza Semântica dos Papéis Temáticos e sua Representação em um Sistema Computacional	
4.4.1 Visão Não-lexicalista dos Papéis Temáticos	
4.4.2 Representações por Microcaracterísticas Semânticas	
4.5 A Composicionalidade	
4.6 Conclusão	

CAPÍTULO 5 - HTRP: UM SISTEMA HÍBRIDO SIMBÓLICO-CONEXIONISTA PARA O PROCESSAMENTO DE PAPÉIS TEMÁTICOS.....71

5.1 Introdução	
5.2 A Saída Erro	
5.3 A Arquitetura Conexionista	
5.4 Representações Baseadas em Microcaracterísticas Semânticas	
5.4.1 O Processador de Papel Temático Híbrido – HTRP	
5.5 As Regras Simbólicas Iniciais para a BIW	
5.6 O Aprendizado	
5.7 As Microcaracterísticas Complementares	
5.8 As Regras Finais	
5.9 Os Verbos do Sistema	

CAPÍTULO 6 - CONCLUSÕES.....86

6.1 Introdução	
6.2 Conclusões e Trabalhos Futuros	

Anexos.....90

ANEXO A - AS REDES NEURAIS ARTIFICIAIS.....91

A.1 Introdução	
A.2 O Neurônio Biológico	

A.2.1	Variantes do Neurônio Clássico	
A.2.2	Sinapses: Junções entre Células Nervosas	
A.2.2.1	As Sinapses são Químicas e não Elétricas	
A.2.2.2	As Sinapses Podem Excitar ou Inibir	
A.2.2.3	Generalizações sobre Sinapses	
A.2.2.4	Peptídeos: Moduladores da Função Sináptica	
A.2.2.5	Peptídeo: Transmissor Lento ou Neuromodulador ?	
A.3	O Cérebro como Modelo	
A.3.1	Paralelismo	
A.3.2	Variedades de Redes Neurais	
A.3.3	Aprendizado Competitivo	
A.3.4	Representações Distribuídas	
A.3.5	Máquinas de Boltzmann	
A.3.6	Processamento de Sentenças	
A.3.7	O Futuro	
A.4	Algoritmos Conexionistas	
A.4.1	Redes Perceptron Multicamadas	
A.4.1.1	O Perceptron	
A.4.1.2	O Algoritmo Backpropagation e a Rede Perceptron Multicamadas	
A.4.1.3	O Algoritmo Backpropagation	
A.4.1.4	Generalização	
A.4.2	Redes Recorrentes	
A.4.2.1	A Representação do Tempo	
A.4.2.2	Conclusões sobre Tarefas Temporais	
A.5	Abordagem Híbrida: As Redes Neurais Baseadas em Conhecimento	
A.6	Conclusão	

ANEXO B - ABORDAGENS AO PROCESSAMENTO SIMBÓLICO DA LINGUAGEM NATURAL.....125

B.1	Introdução
B.2	O Relacionamento entre Regras e Casos
B.3	O Relacionamento entre Regras e Princípios
B.4	<i>Parser</i> Baseado em Princípios
B.5	<i>Parser</i> Baseado em Casos
B.6	Conclusão

ANEXO C - MANUAL DO USUÁRIO DO HTRP.....137

C.1	Introdução
C.2	A Versão RIW
C.3	A Versão BIW
C.4	Desenvolvimento do HTRP
C.4.1	A Arquitetura Conexionista
C.4.2	A “Clusterização” e a “Anulação”
C.4.3	A Otimização
C.4.4	A Rede Recorrente

C.4.5 Os Clusters de Unidades na Saída	
C.4.6 O Gerador de Frases para o Reconhecimento	
C.4.7 As Regras Simbólicas	
C.4.8 As Leituras Alternativas dos Verbos	
C.5 Conclusão	

ANEXO D – GLOSSÁRIO.....	155
--------------------------	-----

SUMMARY.....	165
--------------	-----

REFERÊNCIAS BIBLIOGRÁFICAS.....	166
---------------------------------	-----

BIBLIOGRAFIA CONSULTADA.....	177
------------------------------	-----

Resumo

Em Lingüística, as relações semânticas entre palavras em uma sentença são consideradas, entre outras coisas, através da atribuição de *papéis temáticos*, por exemplo, AGENTE, INSTRUMENTO, etc. Como na lógica de predicados, expressões lingüísticas simples são decompostas em um predicado (frequentemente o verbo) e seus argumentos. O predicado atribui papéis temáticos aos argumentos, tal que cada sentença tem uma *grade temática*, uma estrutura com todos os papéis temáticos atribuídos pelo predicado. Com a finalidade de revelar a grade temática de um sentença semanticamente bem formada, um sistema chamado HTRP (*Hybrid Thematic Role Processor* – Processador de Papéis Temáticos Híbrido) é proposto, no qual a arquitetura conexionista tem, como entrada, uma representação distribuída das palavras de uma sentença, e como saída, sua grade temática. Duas versões do sistema são propostas: uma versão com pesos de conexão iniciais aleatórios – RIW (*random initial weight version*) e uma versão com pesos de conexão iniciais polarizados – BIW (*biased initial weight version*) para considerar sistemas *sem* e *com* conhecimento inicial, respectivamente. Na BIW, os pesos de conexão iniciais refletem regras simbólicas para os papéis temáticos. Para ambas as versões, depois do treinamento supervisionado, um conjunto de regras simbólicas finais é extraído, que é consistentemente correlacionado com o conhecimento lingüístico – simbólico. No caso da BIW, isto corresponde a uma revisão das regras iniciais. Na RIW as regras simbólicas parecem ser induzidas da arquitetura conexionista e do treinamento. O sistema HTRP aprende a reconhecer a grade temática correta para sentenças semanticamente bem formadas do português. Além disso, este sistema possibilita considerações a respeito dos aspectos cognitivos do processamento lingüístico, através das regras simbólicas introduzidas (na BIW) e extraídas (de ambas as versões).

Palavras-chave: Processamento de Linguagem Natural, Redes Neurais, Inteligência Artificial.