

Capítulo 2

Processamento de Linguagem Natural: da Engenharia de Computação à Lingüística

“Não é que eles não possam ver a solução. É que eles não podem ver o problema”

G. K. Chesterton (1874-1936): *Scandal of Father Brown* (1935)

2.1 Introdução

O interesse e a pesquisa em Processamento de Linguagem Natural (PLN) começou no Mestrado, com a dissertação *Redes Neurais e Lógica Formal em Processamento de Linguagem Natural* (Rosa, 1993), defendida em setembro de 1993, junto ao *Departamento de Engenharia de Computação e Automação Industrial* da *Faculdade de Engenharia Elétrica e de Computação* da *Unicamp*, com o Prof. Dr. Márcio Luiz de Andrade Netto, como orientador. As várias abordagens existentes para o PLN se traduzem em aplicação de técnicas de Inteligência Artificial, ou baseadas em lógica ou em sistemas conexionistas. O processador do Mestrado – doravante chamado de P3A (Processador 3 Abordagens: Simbólica, Conexionista e Conexionista-Recorrente) – faz um “mix” destas duas abordagens, acrescentando a extensão temporal da análise da sentença. O resultado é um sistema que faz a análise sintática de sentenças da língua portuguesa, baseada na **lógica de predicados** do **Prolog**, e análises semântica e recorrente, baseadas numa abordagem de redes neurais. Este sistema possui léxico e regras de sintaxe limitados. Mas, para este universo, os resultados obtidos foram bastante satisfatórios, pois o sistema faz a classificação correta em gramatical e não gramatical de sentenças da língua portuguesa, para a grande maioria dos casos (Rosa e Netto, 1994). Mais tarde, um sistema derivado deste – CPPro –, mas já com ênfase em papéis temáticos, foi apresentado, alcançando uma performance muito melhor – para 6000 sentenças semanticamente válidas, o sistema rejeitou apenas 5 e para 3000 inválidas, o sistema aceitou 26 (Rosa, 1997).

O sistema P3A, assim como o sistema de McClelland e Kawamoto (1986), usa o termo **papel de caso** para designar o conceito de *papel de caso temático*, relacionado à noção de caso de Fillmore (1968). O sistema CPPro usa o conceito de *papel temático* da teoria lingüística da Regência e Ligação (Haegeman, 1991). A preocupação com a parte *lingüística* do PLN, considerada “pobre” no P3A, fez com que o Doutorado seguisse outros caminhos. A seguir, apresenta-se um resumo do P3A e do CPPro.

2.2 O Sistema P3A

Existem várias abordagens para o PLN. E existem muitos trabalhos publicados nas vá-

rias abordagens. Entretanto, combinações das diversas abordagens existentes são raras. Este trabalho ousa em combinar uma abordagem sintática baseada na lógica de predicados, uma abordagem conexionista **feedforward**, baseada em atribuições de papéis de caso às palavras, referente à análise semântica, e uma abordagem conexionista recorrente, que leva em consideração características temporais da análise da sentença.

A grande maioria dos trabalhos publicados tratam de processamento da língua inglesa. Houve necessidade da transposição de procedimentos e idéias para a língua portuguesa. E também a necessidade do tratamento da **ambigüidade**, que é o maior desafio enfrentado pelos sistemas que tratam da linguagem natural: identificar o verdadeiro significado de uma determinada palavra pode ser tão complicado, que às vezes só é possível com uma consulta ao usuário. Neste sistema, não se tem a intenção de resolver o problema da ambigüidade, mas apenas contribuir com idéias e apontar direções, quando talvez poder-se-á transpor, ao menos parcialmente, este obstáculo.

2.2.1 A Análise da Forma

O P3A se divide, basicamente, em três etapas. A primeira trata da *análise sintática* de sentenças da língua portuguesa. Esta implementação foi baseada na **Gramática de Cláusulas Definidas** de Pereira e Warren (1980). São sentenças declarativas, compostas por até quatro elementos-chave, que são o sujeito, o verbo, o objeto e o complemento, que pode ser o instrumento ou o modificador. Estes elementos-chave podem vir acompanhados de outras palavras como artigos, adjetivos, partículas reflexivas, etc. A análise sintática faz verificação de concordância de gênero e número, além de montar uma estrutura, chamada de estrutura-chave (contendo apenas os elementos-chave), que alimentará o analisador semântico (segunda etapa). Por exemplo, a sentença

(1) *A menina bonita quebrou a frágil vidraça com um martelo*

gerará a estrutura-chave

(2) *menina-quebrar-vidraça-martelo*

onde *menina* é o sujeito, *quebrar* é o verbo, *vidraça* é o objeto e *martelo* é o instrumento. Note que uma sentença nunca tem, ao mesmo tempo, um instrumento e um modificador.

Uma sentença pode ser

(3) *Todos os homens comem macarrão com cenouras*

em que *cenouras* é o modificador de *macarrão*. A estrutura-chave da sentença anterior será

(4) *homem-comer-macarrão-cenoura*

É claro que uma sentença pode não estar completa. Por exemplo, na sentença

(5) *O homem se moveu*

não há objeto claramente expresso e nem complementos (o verbo *mover* aqui é reflexivo).

O analisador sintático é baseado na lógica de predicados e foi implementado em Prolog. Por exemplo, para analisar a sentença (1), deve haver o seguinte conjunto de cláusulas Prolog de (6) a (20) abaixo¹:

(6) $s \rightarrow sn, sv.$

(7) $sn \rightarrow det, subst.$

(8) $sn \rightarrow det, subst, adj.$

(9) $sn \rightarrow det, adj, subst.$

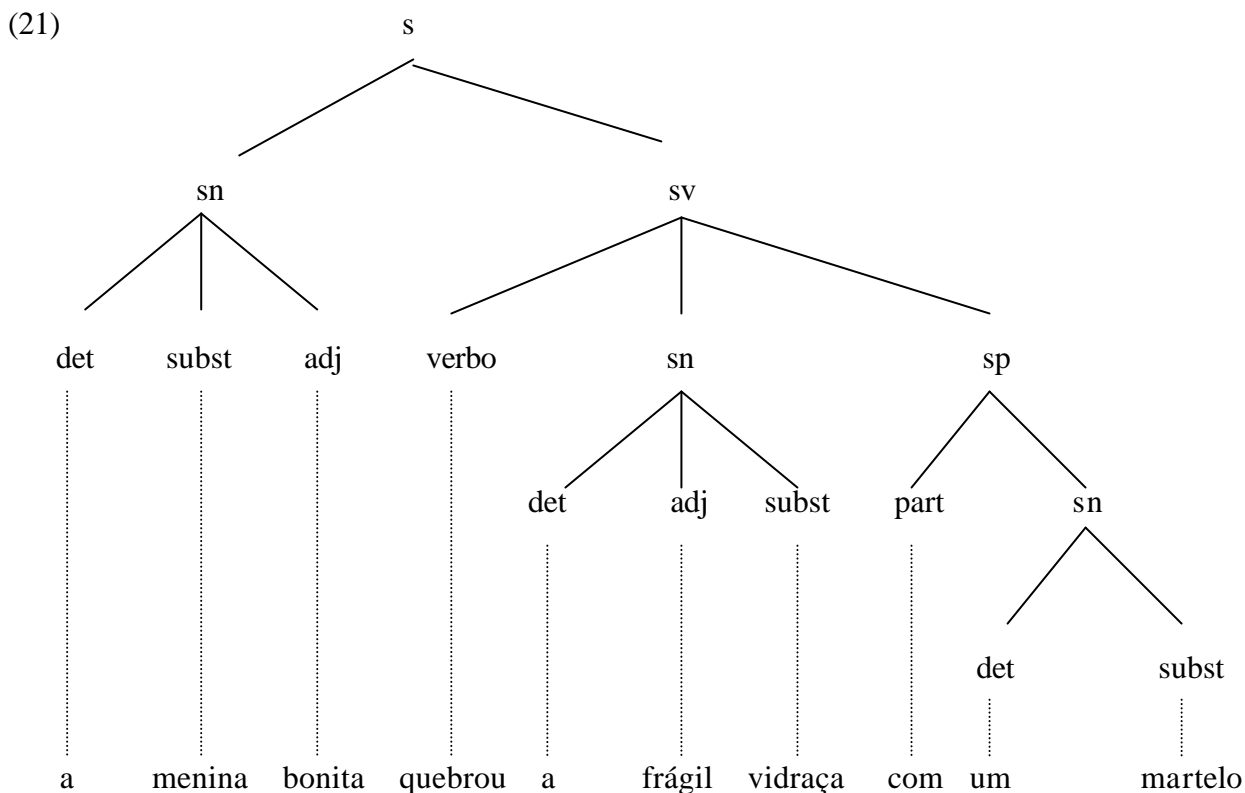
(10) $sv \rightarrow verbo, sn, sp.$

(11) $sp \rightarrow part, sn.$

¹ Note que os predicados em Prolog são representados por letras minúsculas. Nesse caso, se está representando o mesmo que as letras maiúsculas correspondentes (N, V, SN, SV) usualmente representam na literatura lingüística.

- (12) det -- > [a].
- (13) det -- > [um].
- (14) subst -- > [menina].
- (15) subst -- > [vidraça].
- (16) subst -- > [martelo].
- (17) adj -- > [bonita].
- (18) adj -- > [frágil].
- (19) verbo -- > [quebrou].
- (20) part -- > [com].

Por exemplo, a cláusula (6) descreve como toda sentença *s* é formada por um sintagma nominal *sn* seguido de um sintagma verbal *sv*. As cláusulas (7) a (9) dizem que um sintagma nominal pode ser um determinante seguido de um substantivo (7), um determinante, um substantivo e um adjetivo (8) e um determinante, um adjetivo e um substantivo (9). Para a sentença (1), a **árvore de parsing** (21) poderia ser construída.



2.2.2 A Análise do Significado

A segunda etapa trata da *análise semântica*. Nesta etapa a sentença, já analisada sintaticamente pelo analisador sintático e já no formato de estrutura-chave, vai alimentar a entrada de uma rede conexionista perceptron multi-camadas (vide figura 2.6 e Anexo A) com três camadas, que dirá se a sentença tem significado.

Esta etapa foi baseada nos trabalhos de McClelland e Kawamoto (1986) e Waltz e Pollack (1985), que tratam a palavra como um conjunto de microcaracterísticas semânticas. Ou seja, toda palavra é descrita como um vetor de **bits**, onde cada subconjunto de bits tem um significado associado, como *humano-não humano*, *frágil-inquebrável*, *pequeno-médio-grande*, etc. (vide tabelas 2.2 e 2.3, e o item 5.7 – *As Microcaracterísticas Complementares*).

A rede é alimentada não com a sentença propriamente dita, mas com uma representação canônica da estrutura constituinte da sentença (as palavras), ou seja, com o seu conjunto de microcaracterísticas semânticas.

Na verdade, um determinado verbo possui quatro redes a ele associadas: uma para o agente, uma para o paciente, uma para o instrumento e uma para o modificador. Por exemplo, para a sentença (1), a rede do agente é ativada por uma estrutura relacionada às microcaracterísticas de *menina*. Esta estrutura é chamada de *estrutura de sentença* (vide figura 2.4). O formato da saída da rede é chamada de *estrutura de caso* (vide figura 2.5), que é o confronto das microcaracterísticas de *menina* com as do verbo *quebrar*. O processo se repete para as outras redes.

Como as redes foram treinadas para várias sentenças, elas têm condição de verificar se uma sentença nova, ou seja, uma sentença não apresentada anteriormente, está ou não semanticamente correta. O algoritmo usado para implementar estas redes foi o **backpropagation**. Este algoritmo consiste basicamente no seguinte. Primeiro, atribui-se pesos aleatórios às liga-

ções entre os elementos das redes. Ao se entrar com uma estrutura de sentença, a saída da rede é comparada com a saída desejada, ou seja, a saída que deveria ocorrer caso a rede tivesse aprendido aquela estrutura. As ligações são enfraquecidas ou fortalecidas, para “corrigir” a saída da rede. Este processo é repetido dezenas de vezes, até que a rede atinja uma situação de convergência, ou seja, até que a rede “aprenda” aquela estrutura.

O funcionamento do sistema conexionista consiste de duas fases. A primeira, a fase do *aprendizado*, consiste na apresentação seqüencial de sentenças diferentes, mas corretas semanticamente, e na correção de pesos descrita acima, tudo isto dezenas de vezes (**ciclos de treinamento**). A segunda parte, a fase do **reconhecimento**, onde é apresentada uma sentença, nova ou não, à rede, e ela deve ser capaz de, num único ciclo, dizer se esta sentença está correta semanticamente.

Esta etapa foi implementada na linguagem de programação Pascal.

2.2.3 A Análise Temporal

A terceira etapa, a *análise temporal*, trata da verificação das seqüências de palavras previstas. Pode-se, com esta etapa, verificar se as palavras estão numa seqüência apropriada. Assim, verifica-se mais uma vez, a *ordem* das palavras constituintes de uma sentença (sintaxe) e também as **restrições de seleção**, ou seja, quais palavras podem ocupar quais posições argumentais (semântica). Esta abordagem foi baseada no trabalho de Elman (1990).

A análise semântica não faz verificação, em uma sentença, dos relacionamentos entre todos os elementos. Só entre cada elemento e o verbo e entre o objeto e o modificador, no caso de este existir. Ou seja, na sentença (3), apenas existe verificação de relação entre *homem* e *comer*, entre *macarrão* e *comer* e entre *macarrão* e *cenoura*. Não há verificação entre *homem* e *macarrão* e nem entre *homem* e *cenoura*. Caso se deseje fazer esta verificação, utiliza-se uma **rede recorrente**, que é uma rede provida de memória, onde pode-se conhecer uma determinada seqüência de palavras que se ensinou (veja Anexo A – As Redes Neurais Artificiais).

O algoritmo empregado nesta rede também foi o *backpropagation*. A diferença é que esta rede tem uma camada a mais, onde se armazena o estado anterior (memória). Na fase do aprendizado, ensina-se à rede todas as seqüências de palavras possíveis para todas as sentenças possíveis. Na fase de reconhecimento, entra-se com uma palavra e a rede fornece a próxima na seqüência. Esta etapa foi implementada conjuntamente com o analisador semântico, portanto também em Pascal.

2.2.4 Arquitetura do Modelo

A meta primária deste modelo é prover um mecanismo que possa começar a considerar conjuntamente o papel da ordem da palavra e das restrições semânticas na atribuição do papel de caso. Deseja-se que o modelo seja capaz de *aprender* a fazer isto baseado em experiência com sentenças e suas representações de caso. Deseja-se que este modelo seja capaz de *generalizar* o que aprendeu para novas sentenças formadas de combinações de palavras.

O modelo consiste de dois conjuntos de unidades: um para representar a estrutura superficial da sentença e um para representar sua estrutura de caso². O modelo aprende através de apresentações de pares corretos de estrutura superficial e estrutura de caso; durante o **treinamento**, é apresentada à entrada a estrutura superficial e examina-se a saída que o modelo gera no nível de estrutura de caso.

As sentenças processadas pelo modelo consistem de um verbo e de um a três sintagmas nominais (SNs). Existe sempre um SN Sujeito e opcionalmente pode existir um SN Objeto. Se este estiver presente, pode também existir um com-SN; isto é, um SN em uma sintagma preposicional de final de sentença começando com a palavra *com*. As limitações de tipos de sentenças tratadas são necessárias por razões práticas, de implementação. Mas isso não desmerece o modelamento, que visa, antes de mais nada, começar a estudar o problema da atribuição de papéis temáticos, e não pretende apresentar uma “máquina” pronta e nem

² McClelland e Kawamoto (1986) chamam de *estrutura de caso* o tipo de representação distribuída que relaciona o substantivo com o verbo em determinada sentença. É a saída do sistema conexionista (veja figura 2.5).

mesmo uma teoria pronta.

2.2.5 Microcaracterísticas Semânticas

Nos formatos de entrada canônicos, as palavras são representadas como listas de microcaracterísticas semânticas (Waltz e Pollack, 1985; McClelland e Kawamoto, 1986). Para substantivos e verbos, as características são agrupadas em muitas dimensões. Cada dimensão consiste de um conjunto de valores mutuamente exclusivos e, assim, cada palavra não ambígua é representada por um vetor no qual um, e apenas um, valor em cada dimensão está **ativo** para a palavra e todos os outros valores estão **desativados**. Valores que estão ativos são representados nos vetores de características como “1”s. Valores que estão desativados são representados por pontos (“.”).

Foram escolhidas as dimensões e os valores em cada dimensão para capturar o que se considerou aspectos importantes de variações semânticas nos significados de palavras que tinham implicações para a atribuição de papéis das palavras (McClelland e Kawamoto, 1986).

O conjunto completo das dimensões usadas nos conjuntos de características é dado na tabela 2.1 (adaptada de McClelland e Kawamoto, 1986). Algumas dimensões dos substantivos merecem comentário: *softness* representa dureza (*hard*, como em *escrivainha*) e não rígido (*soft*, como em *menina*); quanto ao *volume*, obviamente é relativo às palavras pertencentes ao léxico adotado; quanto à *forma*, *compacto* representa coisas muito pequenas, como *garfo*; *2-D* representa coisas “planas”, como *cortina* e *3-D*, coisas tridimensionais, como *homem*; *ponta*, com os valores *pontiagudo*, que representa palavras “com ponta”, como *garfo*, e *redondo*, que representa coisas “arredondadas”, como *bola*.

Em relação às dimensões dos verbos, há a necessidade de algum esclarecimento. A dimensão *realizador* indica se existe um Agente instigando o evento. A dimensão *causa* especifica se o verbo é causal. Se não, é porque não existe causa especificada (como no caso de *a vidraça quebrou*) ou é porque não há troca (como no caso de *o garoto tocou a menina*). A dimensão *toque* indica se o Agente, o Instrumento, ambos ou nenhum toca o Paciente. A di-

mensão *nat_troca* especifica a natureza da troca que tem lugar no Paciente. *agt_mvmt* e *pt_mvmt* especificam o movimento do Agente e do Paciente, respectivamente; e *intensidade* simplesmente indica a força da ação. Os rótulos dados às dimensões são, é claro, apenas para referência; eles são escolhidos de tal forma que cada dimensão de substantivo ou verbo tenha uma única primeira letra que possa ser usada para designar a dimensão (veja figuras 2.4 e 2.5). Deve-se enfatizar que outras características podem ser incluídas para estender o modelo para conjuntos maiores de substantivos e verbos.

<i>SUBSTANTIVOS – DIMENSÕES</i>	<i>VALORES DE CARACTERÍSTICAS</i>
HUMANO (HU)	humano, não humano
<i>SOFTNESS</i> (SO)	<i>soft, hard</i>
GÊNERO (GE)	masculino, feminino
VOLUME (VOL)	pequeno, médio, grande
FORMA (FOR)	compacto, 2-D, 3-D
PONTA (PO)	pontiagudo, redondo
DUREZA (DU)	frágil, inquebrável
TIPO-OBJ (TIP)	alimento, brinquedo, ferramenta/utensílio, animado
<i>VERBOS – DIMENSÕES</i>	<i>VALORES DE CARACTERÍSTICAS</i>
REALIZADOR (RE)	sim, não
CAUSA (CA)	sim, não
TOQUE (TOQU)	agente, instrumento, ambos, nenhum
NAT_TROCA (N_TR)	peças, fragmentos, química, nenhum
AGT_MVMT (A_M)	transformador, parcial, nenhum
PT_MVMT (P_M)	transformador, parcial, nenhum
INTENSIDADE (IN)	baixo, alto

Tabela 2.1. Dimensões de verbos e substantivos e seus valores de características.

As tabelas 2.2 e 2.3 dão os vetores que se atribui a algumas das palavras usadas no modelo (adaptadas de McClelland e Kawamoto, 1986).

Um das metas para o modelo é mostrar como ele pode selecionar o significado apropriado no contexto para uma palavra ambígua. Para palavras ambíguas (*galinha*, viva ou cozida) o padrão de entrada é a média dos padrões de características de cada uma das duas leituras da palavra. Isto significa que nos casos onde as duas concordam com o valor de uma dimensão de entrada particular, esta dimensão tem o valor acordado na representação de entrada. Nos casos onde os dois discordam, a característica tem o valor de .5 (representado por “?”) na representação de entrada. Uma meta é ver se o modelo pode corretamente preencher

estes valores não especificados, efetivamente recuperando os valores perdidos do contexto no processo da atribuição da palavra ao papel de caso apropriado. A tabela 2.2 indica as duas leituras de *galinha*, assim como as formas ambíguas usadas como entradas³.

SUBSTANTIVOS	HU	SO	GE	VOL	FOR	PO	DU	TIP_
alimento	.1	1.	1.	1..	???	.1	1.	1...
boneca	.1	1.	.1	1..	..1	.1	1.	.1..
galinha	.1	1.	.1	1..	..?	.1	??	?..?
galinha cozida	.1	1.	.1	1..	1..	.1	1.	1...
galinha viva	.1	1.	.1	1..	..1	.1	1.	...1
garoto	1.	1.	1.	.1.	..1	.1	.1	...1
homem	1.	1.	1.	..1	..1	.1	.1	...1
macaco	.1	??	1.	1..	..1	1.	.1	..??
macaco ferram.	.1	.1	1.	1..	..1	1.	.1	..1.
macaco animal	.1	1.	1.	1..	..1	1.	.1	...1
mulher	1.	1.	.1	..1	..1	.1	.1	...1
pedra	.1	.1	.1	1..	..1	1.	.1	..1.
vidraça	.1	.1	.1	.1.	.1.	1.	1.	..1.

Tabela 2.2. Alguns substantivos usados no modelo e suas microcaracterísticas semânticas.

	RE	CA	TOQU	N_TR	A_M	P_M	IN
bateu	1.	.1	.1..	...1	.1.	..1	.1
bateuAVPI	1.	.1	.1..	...1	.1.	..1	.1
bateuAVP	1.	.1	1...	...1	.1.	..1	.1
bateuIVP	.1	.1	.1..	...1	..1	..1	.1

Nota: As formas verbais, seguidas por cadeias de letras maiúsculas representam os padrões de características alternativas entre os quais o modelo pode escolher, para especificar a leitura apropriada do verbo no contexto. Estes padrões de características alternativas correspondem às características semânticas do verbo apropriado para configurações particulares de papéis de caso, como indicado pelas letras: A = Agente, V = Verbo, P = Paciente, I = Instrumento. A posição da letra indica a posição do constituinte correspondente na sentença de entrada. Os padrões dados com o verbo genérico (sem as letras) são usados nas representações de entrada, do nível de sentença. Além deste verbo, estão previstos os verbos *comer*, *mover* e *quebrar*.

Tabela 2.3. Um dos verbos usados no modelo e suas representações de microcaracterísticas

³ Para o conceito *alimento* que é tido como o paciente implícito em sentenças como *o garoto comeu*, nenhuma forma particular parece apropriada. Portanto, a representação de saída pretendida é assumida como não especificada (indicada por “?”) para todos os valores da dimensão “forma”. Para todas as outras dimensões, *alimento* tem os valores típicos para alimentos.

2.2.6 Unidades de Estrutura de Sentença.

Baseado em McClelland e Kawamoto (1986), a representação do nível de estrutura de sentença de uma sentença de entrada não é o conjunto de vetores de microcaracterísticas dos constituintes; ela é o padrão de ativação que esses vetores produzem sobre as unidades que correspondem a *pares* de características. Estas unidades são chamadas unidades de estrutura de sentença (ES). A figura 2.4 (Rosa, 1993) mostra estruturas de sentença para a sentença (22). A linha superior mostra o vetor de microcaracterísticas dos constituintes da sentença (22), com o símbolo “.” (ponto) representando o valor *desativado*, o valor “1” representando o valor *ativo* e o valor “?” representando o valor *indeterminado*. As letras abaixo do vetor e ao lado da matriz representam as dimensões de cada palavra (veja tabela 2.2). A estrutura de sentença contrasta as microcaracterísticas de cada constituinte com elas mesmas.

(22) O garoto quebrou a vidraça com o martelo.

Cada unidade ES representa a conjunção de duas microcaracterísticas de uma palavra da sentença. Como há quatro itens lexicais na estrutura de uma sentença (verbo, sujeito, objeto e complemento), existem quatro conjuntos de unidades ES. Dentro de cada conjunto existe uma unidade que representa a conjunção de todo valor de microcaracterística em cada dimensão com todo valor de microcaracterística em qualquer outra dimensão.

Embora o modelo funcione bem com esta simulação, presume-se que simulações que usem um léxico maior requereria maior diferenciação de algumas representações de substantivos e verbos.

2.2.7 Representação de Papel de Caso

A representação de papel de caso tem uma forma levemente diferente da representação de estrutura de sentença. Para entender esta representação, é útil voltar a um ponto de vista mais abstrato e considerar mais genericamente como se deve representar uma descrição es-

trutural numa **representação distribuída**. Em geral uma descrição estrutural pode ser representada por um conjunto de triplas da forma (A R B) onde A e B correspondem aos nós na descrição estrutural e R representa a relação entre os nós. Por exemplo, uma hierarquia de inclusão de classes pode ser representada por triplas da forma (X *é-um* Y), onde X e Y são nomes de categorias. Qualquer outra descrição estrutural, seja uma estrutura de constituinte sintático, uma estrutura de constituinte semântico ou qualquer outra coisa, pode ser representada desta forma. Especificamente, a atribuição de papel de caso dos constituintes da sentença *O garoto quebrou a vidraça com o martelo* pode ser representada como em (23) (McClelland e Kawamoto, 1986).

- (23) *Quebrou* Agente Garoto
 Quebrou Paciente Vidraça
 Quebrou Instrumento Martelo

A estrutura constituinte de uma sentença tal como *O garoto comeu o macarrão com molho* poderia ser representada por (24).

- (24) *Comeu* Agente Garoto
 Comeu Paciente Macarrão
 Macarrão Modificador Molho

As estruturas de caso (saída da rede conexionista) para as relações (23) estão representadas na figura 2.5 (Rosa, 1993). Da mesma forma que na figura 2.4, as letras minúsculas representam as dimensões dos substantivos e verbo envolvidos na sentença (22). Observe que, no caso das estruturas de caso, há o contraste das microcaracterísticas de cada substantivo com o verbo *quebrar*.


```

1.1..1..1....1...1.1
rrccttttnnnnaaapppii
r
r
c 1?
c ?.
t ?.?.
t 1?1?
t ?.?.
t ?.?.
n 1?1??1??
n ?.?.?..
n ?.?.?..
n ?.?.?..
a ?.?.?..?....
a 1?1??1??1??
a ?.?.?..?....
p ?.?.?..?....?.
p ?.?.?..?....?.
p 1?1??1??1??1??1?
i ?.?.?..?....?....?
i 1?1??1??1??1??1??1

```

Figura 2.4A. Estrutura de sentença para *quebrou*.

```

.1.1.1.1..1.1.1...1.
hhssggvvvfffpddtttt
h
h
s .?
s ?1
g .??.
g ?1?1
v .??.?.
v ?1?1?1
v .??.?.
f .??.?..?.
f ?1?1?1?1?1?
f .??.?..?.
p ?1?1?1?1?1?1?
p .??.?..?..?.
d ?1?1?1?1?1?1?1?
d .??.?..?..?..?.
t .??.?..?..?..?..?.
t .??.?..?..?..?..?.
t ?1?1?1?1?1?1?1?1?
t .??.?..?..?..?..?.

```

Figura 2.4C. Estrutura de sentença para *vidraça*.

```

1.1.1..1...1.1.1...1
hhssggvvvfffpddtttt
h
h
s 1?
s ?.
g 1?1?
g ?.?.
v ?.?.?.
v 1?1?1?
v ?.?.?.
f .?..?..?.
f ?.?.?..?..
f 1?1?1?1?1?
p ?.?.?..?..?..
p 1?1?1?1?1?1?1?
d ?.?.?..?..?..?..
d 1?1?1?1?1?1?1?1?
t ?.?.?..?..?..?..?
t ?.?.?..?..?..?..?
t 1?1?1?1?1?1?1?1?1
t ?.?.?..?..?..?..?

```

Figura 2.4C. Estrutura de sentença para *garoto*.

```

.1.11.1..1...1.1...1.
hhssggvvvfffpddtttt
h
h
s .?
s ?1
g ?1?1
g .??.
v ?1?11?
v .?..?..
v .?..?..
f .?..?..?..
f ?1?11?1?1?
f .?..?..?..
p .?..?..?..?..
p ?1?11?1?1?1?1?
d .?..?..?..?..?..
d ?1?11?1?1?1?1?1?
t .?..?..?..?..?..?
t .?..?..?..?..?..?
t ?1?11?1?1?1?1?1?1
t .?..?..?..?..?..?

```

Figura 2.4D. Estrutura de sentença para *martelo*.


```

hhssggvvvfffpddtttt
r 1.1.1..1...1.1.1..1.
r .....
c 1.1.1..1...1.1.1..1.
c .....
t .....
t 1.1.1..1...1.1.1..1.
t .....
t .....
n 1.1.1..1...1.1.1..1.
n .....
n .....
n .....
a .....
a 1.1.1..1...1.1.1..1.
a .....
p .....
p .....
p 1.1.1..1...1.1.1..1.
i .....
i 1.1.1..1...1.1.1..1.

```

Figura 2.5A. Estrutura de caso para *garoto-quebrou*

```

hhssggvvvfffpddtttt
r .1.11.1...1..1.1..1.
r .....
c .1.11.1...1..1.1..1.
c .....
t .....
t .1.11.1...1..1.1..1.
t .....
t .....
n .1.11.1...1..1.1..1.
n .....
n .....
n .....
a .....
a .1.11.1...1..1.1..1.
a .....
p .....
p .....
p .1.11.1...1..1.1..1.
i .....
i .1.11.1...1..1.1..1.

```

Figura 2.5C. Estrutura de caso para *quebrou-martelo*

```

hhssggvvvfffpddtttt
r .1.1.1.1..1.1.1...1.
r .....
c .1.1.1.1..1.1.1...1.
c .....
t .....
t .1.1.1.1..1.1.1...1.
t .....
t .....
n .1.1.1.1..1.1.1...1.
n .....
n .....
n .....
a .....
a .1.1.1.1..1.1.1...1.
a .....
p .....
p .....
p .1.1.1.1..1.1.1...1.
i .....
i .1.1.1.1..1.1.1...1.

```

Figura 2.5B. Estrutura de caso para *quebrou-vidraça*

2.2.8 Detalhes do Processamento de Sentença e Aprendizado

Considere a figura 2.6. Na apresentação de uma sentença, a entrada da rede para cada uma das unidades de estrutura da sentença é determinada, com base nos vetores de características das palavras (corresponde à camada de entrada, aos valores x_i , $1 < i < A$). Cada unidade de estrutura de superfície tem conexões com peso associado modificável ($w1_{ij}$ e $w2_{ij}$) para cada uma das unidades de estrutura de caso (corresponde à camada de saída, valores o_k , $1 < k < C$), através da camada escondida (h_j , $1 < j < B$) e cada unidade de estrutura de caso tem uma polarização modificável (equivalente a uma conexão a partir de uma unidade especial que está sempre ligada). Baseada no padrão de estrutura de sentença e nos valores correntes dos pesos, uma entrada da rede para cada unidade de estrutura de caso é computada. Unidades de estrutura de caso têm valores de ativação 0 e 1 e a ativação é uma função da entrada da rede, que é implementada com o algoritmo *backpropagation*.

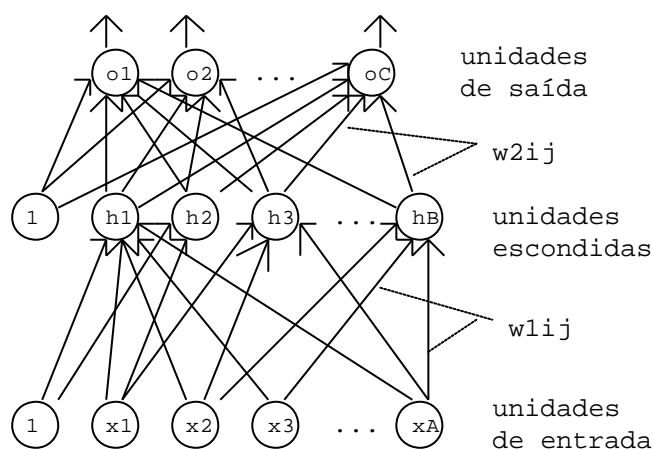


Figura 2.6. Uma rede perceptron multicamada (Rich e Knight, 1994).

Durante o aprendizado, a ativação resultante de cada unidade de estrutura de caso é comparada ao valor que ela deveria ter na leitura correta da sentença. A leitura correta é fornecida como uma “entrada mestre” especificando quais unidades de papéis de caso devem estar ligadas. O aprendizado simplesmente corresponde ao ajuste de pesos de conexão para

fazer a saída gerada pelo modelo corresponder da forma mais próxima possível à entrada mestre.

2.2.9 Experimentos de Simulação

O mais importante sobre o modelo é o fato de que sua resposta a novas entradas é estritamente dependente de sua experiência. Na avaliação de seu comportamento então, é importante conhecer ao que ele foi exposto durante o aprendizado.

A experiência principal consiste na geração de várias sentenças derivadas dos **frames** de sentenças listados na tabela 2.7 (adaptada de McClelland e Kawamoto, 1986). Deve ser enfatizado que estes frames de sentenças são simplesmente usados para gerar um conjunto de sentenças válidas. Cada *frame* especifica um verbo, um conjunto de papéis e uma lista de possíveis preenchedores de cada papel. Portanto, o *frame* de sentença *O humano quebrou o objeto_frágil com o quebrador* é simplesmente um gerador para todas as sentenças na qual *humano* é substituído por uma das palavras na lista de humanos na tabela 2.8 (adaptada de McClelland e Kawamoto, 1986), *objeto_frágil* é substituído por uma das palavras na lista de objetos frágeis na tabela 2.8 e *quebrador* é substituído por uma das palavras da lista de quebradores na tabela. É claro que estes geradores não capturam todas as propriedades dos elementos envolvidos em cenários reais e então não se pode esperar que o modelo represente fielmente todas as sutilezas.

FRAMES DE SENTENÇA	ATRIBUIÇÃO DE ARGUMENTOS
O HUMANO comeu.	AVF
O HUMANO comeu o ALIMENTO.	AVP
O HUMANO comeu o ALIMENTO com o ALIMENTO.	AVPM
O HUMANO comeu o ALIMENTO com o UTENSÍLIO.	AVPI
O ANIMAL comeu.	AVF
O PREDADOR comeu a PRESA.	AVP

Nota: Atribuições de argumentos especificam a atribuição de papel de caso dos constituintes de uma sentença da esquerda para a direita. A = Agente, V = Verbo, P = Paciente, I = Instrumento, M = Modificador, F = Alimento (implícito), S = Reflexivo (“Self”).

Tabela 2.7. Geradores para sentenças usadas no treinamento e testes.

Para o experimento principal, a tabela 2.8 foi implementada apenas com duas palavras de cada classe para alimentar o gerador automático de sentenças. Isto para permitir que o treinamento da rede se dê apenas com algumas sentenças possíveis, as sentenças de testes *familiares*. Mas foram implementadas todas as palavras da tabela 2.7, ou seja, todas as outras sentenças possíveis, previstas pelas tabelas 2.7 e 2.8 são consideradas sentenças *novas*. Estas sentenças não foram usadas para treinar o modelo⁴.

Ao modelo foi dado 20 ciclos de treinamento, com o conjunto de sentenças de treino. Em cada ciclo, cada sentença era apresentada, a resposta do modelo era gerada e os pesos de conexão era ajustados de acordo com o procedimento *backpropagation*.

humano	homem mulher garoto
animal	macaco-an galinha-vi
objeto	macaco-me galinha-co boneca pedra vidraça
coisa	humano animal objeto
predador	
presa	galinha-vi
alimento	galinha-co
utensílio	
objeto_frágil	vidraça
batedor	macaco-me pedra
quebrador	macaco-me pedra
propriedade	macaco-me boneca

Nota: macaco-an = macaco animal; macaco-me = macaco mecânico (instrumento para trocar pneus) e galinha-vi = galinha viva (animal); galinha-co = galinha cozida (alimento).

Tabela 2.8. Categorias de substantivos

2.2.10 Conclusão

O sistema P3A trouxe como contribuição ao PLN, um sistema com duas etapas: a primeira, com o processamento de sentenças do português, através de uma abordagem lógica, ou

seja, um sistema simbólico baseado em regras da Gramática de Cláusulas Definidas; e uma segunda etapa, conexionista, através de duas redes neurais, a primeira uma rede *feedforward*, onde sentenças, já analisadas previamente em relação à forma, são analisadas do ponto de vista semântico, e a segunda uma rede recorrente, onde são verificadas seqüências de palavras válidas para esta análise.

A abordagem lógica (simbólica) realiza a análise sintática, ou seja, dada uma **gramática sintagmática** (livre de contexto), o sistema, através da técnica **bottom-up** e **backward**, verifica se a sentença está correta sintaticamente. Se estiver, o sistema gera uma estrutura, chamada de estrutura-chave, sem determinantes, adjetivos, etc., que servirá de entrada para a análise conexionista (*rede feedforward* e recorrente). Na rede *feedforward*, o sistema verifica a semântica da sentença, ou seja, se a sentença sintaticamente correta também tem significado. Já a análise temporal serve apenas para verificar a possível seqüência esperada de palavras (ou seja, quais são as palavras esperadas na seqüência da sentença).

O desempenho alcançado pelo sistema foi considerado muito bom, pois a grande maioria das sentenças verificadas, dentro do léxico considerado, foi analisada corretamente (100% análise sintática e 95% análise conexionista *feedforward* e recorrente).

2.3 O Sistema CPro

No sistema CPro – *Connectionist Portuguese Language Processor* (Rosa, 1997), apresenta-se uma adaptação dos modelos conexionistas de McClelland e Kawamoto (1986) e Waltz e Pollack (1985) de representações de microcaracterísticas semânticas das palavras. Neste sistema, que tem como origem o sistema P3A, as microcaracterísticas são baseadas nos relacionamentos temáticos entre as palavras em sentenças do português.

⁴ Alguns geradores (por exemplo, *o humano bateu na coisa com o batedor*) geram um grande número de sentenças diferentes, mas outros geram poucas sentenças. Por esta razão, limitou-se em duas palavras para cada tipo de sentença (tabela 2.8).

2.3.1 As Metas do CPPro

A principal meta do CPPro é prover um mecanismo que lide com a função das restrições semânticas na atribuição de papel temático⁵. O modelo deve ser capaz de aprender a fazer isto baseado na experiência com sentenças e suas representações temáticas, e deve ser capaz de generalizar para novas sentenças.

As entradas do modelo não são as sentenças mas sim representações de microcaracterísticas semânticas das estruturas constituintes das sentenças, como em Rosa e Netto (1994). Para substantivos, vale as mesmas dimensões do P3A já especificadas na tabela 2.1, mas para os verbos, há mudanças (veja tabela 2.9). Como no P3A, as características são agrupadas em várias dimensões, onde cada dimensão consiste de um conjunto de valores mutuamente exclusivos. Cada palavra é representada por um vetor de 20 bits no qual um, e apenas um valor em cada dimensão está *ativo* e todos os outros *desativados*. Valores que estão ativos são representados nos vetores de características como “1”s. Valores que estão desativados são representados como “0”s (veja tabela 2.9 e tabela 2.10).

VERBOS – DIMENSÕES	VALORES TEMÁTICOS
AGENTE (4 bits)	animado, inanimado, experienciador, nenhum
PACIENTE (4 bits)	animado, inanimado, tema, nenhum
INSTRUMENTO (2 bits)	tem, não tem
TOQUE (4 bits)	agente, tema, ambos, nenhum
BENEFÍCIO (2 bits)	sim, não
LOCALIDADE (4 bits)	fonte, meta, locação, nenhum

Tabela 2.9. Papéis temáticos atribuídos por verbos (baseado na classificação de papel temático de Haegeman (1991) e Dowty (1989)).

⁵ Observe que, no sistema CPPro, utiliza-se o termo mais atual *papel temático* em vez de *papel de caso* usado no P3A.

VERBO	<i>agente</i>	<i>paciente</i>	<i>instrumento</i>	<i>toque</i>	<i>benefício</i>	<i>localidade</i>
AMAR	0010	0010	01	0001	01	0001
BATER	1000	0100	10	0100	01	0001
COMER	1000	0100	10	0100	01	0001
DAR	1000	0010	01	0001	10	0100
MOVER	1000	0001	01	0001	01	1000
QUEBRAR	1000	0100	10	0100	01	0001
VER	0010	0010	01	0001	01	0010

Tabela 2.10. Vetores de microcaracterísticas temáticas de alguns verbos. Veja Tabela 2.9.

2.3.2 Experimentos de Simulação

A rede usada no CPPro tem três camadas: a **camada de entrada**, para o qual a estrutura de entrada está disponível; a **camada escondida**, que permite que a rede desenvolva representações internas; e a **camada de saída**, a partir da qual a representação de saída é gerada pelo modelo.

As sentenças apresentadas à rede são geradas preenchendo cada posição de categoria dos *frames* de sentenças. Cada *frame* especifica um verbo, um conjunto de papéis temáticos e uma lista de possíveis preenchedores de cada papel temático. Tal que, o *frame* de sentença *o humano deu o objeto para o humano* é um gerador para sentenças nas quais *humano*, o agente e o beneficiário, são substituídos por uma das palavras na lista de humanos, como *menino* ou *homem*, e *objeto* (o tema) é substituído por uma das palavras na lista de objetos, como *macaco-me* (macaco mecânico), já que *deu* pede por um agente (aquele que dá), um tema (a coisa dada) e um beneficiário (a pessoa que recebe a coisa). Então a sentença *o menino deu o macaco-me para o homem* pode ser gerada. Cada verbo tem o seu gerador. Veja a tabela 2.11 para o verbo *comer*. Note que nos dois últimos *frames* de sentenças não há objeto.

FRAME DE SENTENÇAS	PAPÉIS TEMÁTICOS
O humano comeu o alimento com o utensílio.	agente animado – paciente inanimado – instrumento
O humano comeu o alimento.	agente animado – paciente inanimado
O humano comeu.	agente animado
O animal comeu.	agente animado

Tabela 2.11. O gerador para algumas sentenças com o verbo COMER. Veja a Tabela 2.12.

CATEGORIA	ALGUNS PREENCHEDORES
HUMANO	homem, menina
ALIMENTO	batata, frango, queijo
UTENSÍLIO	colher, garfo
ANIMAL	macaco-an

Tabela 2.12. Algumas categorias de substantivos para alguns preenchedores.

2.3.3 Conclusão

CPPro trouxe, como uma contribuição ao PLN, uma abordagem conexionista à língua portuguesa. Este sistema trabalha com relacionamentos temáticos entre as palavras em uma sentença. CPPro trata de algumas particularidades da língua portuguesa, como a ausência do sujeito ou do objeto.

O sistema pioneiro de McClelland e Kawamoto (1986) trata padrões de relacionamentos, padrões estes que são representações entre palavras de uma sentença. No sistema CPPro, uma arquitetura conexionista baseada em uma adaptação deste modelo é proposta. A representação das palavras é feita através de *vetores de microcaracterísticas semânticas*, formado por subconjuntos que representam aspectos do significado das palavras, como por exemplo, *humano* e *não humano*, onde apenas um valor em cada subconjunto está ativado. No caso do verbo, estes vetores são arranjados na base dos relacionamentos temáticos entre o verbo e as outras palavras de uma sentença, isto é, o modelo pretende mapear papéis temáticos em características semânticas. O objetivo do CPPro é empregar a idéia da representação de microcaracterística com a finalidade de construir uma arquitetura capaz de analisar e aprender as atribuições de relacionamentos temáticos corretas das palavras em uma sentença. A saída do CPPro reflete julgamentos de aceitabilidade semântica de uma determinada sentença. No entanto, a crítica comum aos sistemas conexionistas, a saber, que o sistema funciona mas sem que se conheçam suas determinações internas, aplica-se ao CPPro. Em outras palavras, um sistema que aprenda a reconhecer a associação sistemática de papéis temáticos a sentenças *não* explica o mecanismo interno de atribuição temática.

2.4 Considerações sobre os Sistemas

Os sistemas P3A (Rosa, 1993) e CPPro (Rosa, 1997), apesar de produzirem resultados bem interessantes para o que foram projetados, são sistemas de PLN lingüisticamente empobrecidos, ou seja, teorias lingüísticas ou psicolingüísticas ingênuas são utilizadas nestes sistemas. E além disso, alguns fatores os tornam lingüisticamente redundantes (supérfluos). Veja, por exemplo, o P3A. Este sistema contém uma rede recorrente que faz a análise da sequência de palavras em uma sentença. Pode-se dizer que a análise temporal, para esta finalidade lingüística, pode ser descartada, pois a rede recorrente, como é aplicada no P3A, é substituída parcialmente pela análise sintática. E a análise sintática é uma análise da ordem, pelo menos nas Gramáticas de Cláusulas Definidas. Substitui também, as restrições de seleção, ou seja, só determinadas palavras podem assumir determinadas posições na sentença, certamente um outro tipo de análise da ordem. Na verdade, a rede recorrente aprende qual deve ser a próxima palavra em uma sequência e desta forma, restringe as palavras que podem ocupar determinados lugares.

O sistema CPPro já incorpora uma evolução para um sistema lingüisticamente plausível, com a tentativa de tratamento de papéis temáticos de forma mais enfática. Mas continua com muitas fraquezas lingüísticas. CPPro ainda não se baseia em nenhuma teoria lingüística, apesar de já eliminar a análise temporal. Outra vantagem em relação ao P3A foi a retirada da chamada “análise sintática” do CPPro, realizada com a gramática de cláusulas definidas (GCD) de Pereira e Warren (1980), considerada ultrapassada. Mais uma vantagem do CPPro foi a inclusão de particularidades da língua portuguesa, como por exemplo, sujeito e objeto nulo, ausentes no P3A. O sistema CPPro julga a aceitabilidade semântica de sentenças da língua portuguesa, baseado apenas na representação de microcaracterísticas semânticas das palavras e numa abordagem conexionista, utilizando teorias lingüísticas ingênuas.

2.5 Conclusão Geral

Tanto o sistema P3A quanto o sistema CPro trouxeram como contribuição para o PLN, uma abordagem mista de lógica e conexionismo. Alcançou-se, com os mesmos, resultados bastante satisfatórios dentro do plano proposto. A idéia é continuar estes trabalhos, propondo uma versão baseada em teorias lingüísticas, com vocabulário e estruturas mais ricos, em que se pretende que formações mais complexas da língua portuguesa possam ser tratados. Para isto é necessário adotar uma teoria lingüística para os papéis temáticos. É necessário também, alterar o analisador semântico, construindo redes maiores, onde se permita adequar o tamanho dos vetores de microcaracterísticas semânticas às novas dimensões dadas às palavras, absolutamente necessárias para estabelecer uma diferenciação entre as mesmas. Pode-se, utilizando máquinas mais rápidas, aumentar o número de ciclos para treinamento das redes, contribuindo com isso para uma maior eficiência. E, para garantir uma performance maior, pode-se trabalhar com os sistema híbridos, que são sistemas que combinam a abordagem simbólica (lógica) ao PLN com a abordagem conexionista. O sistema HTRP – *Hybrid Thematic Role Processor* ou Processador de Papel Temático Híbrido – é uma evolução natural destes dois sistemas, com a preocupação de incorporar uma teoria lingüística, a teoria dos papéis temáticos. E de implementar a abordagem híbrida simbólico-conexionista, com a finalidade de aliar as vantagens da abordagem simbólica com o conexionismo.