4 UMA ARQUITETURA ADAPTATIVA PARA PROCESSAMENTO DE LINGUAGEM NATURAL

Este capítulo apresenta uma proposta de arquitetura para um software destinado a efetuar o processamento – mais especificamente, a análise – de linguagem natural, empregando para isso técnicas de análise que fazem uso da tecnologia adaptativa.

Para isso, na proposta descrita nesta tese, a representação externa da linguagem natural se faz com o emprego de uma meta-linguagem, a qual é transformada em um dispositivo reconhecedor que opera como um transdutor baseado em autômatos de pilha estruturados adaptativos.

Esse transdutor opera em quatro níveis sobre o texto de entrada:

- a extração de palavras do texto, sua classificação segundo a categoria a que pertence, sua decomposição em partículas elementares e sua etiquetação ficam a cargo de um autômato adaptativo;
- a parte correspondente à verificação sintática das regras de colocação fica a cargo de outro autômato adaptativo, encarregado da verificação da parte estática da sintaxe da linguagem. Embora neste nível seja, a rigor, desnecessário empregar autômatos adaptativos, sua utilização facilita muito a verificação sintática de textos cujos componentes estejam elipticamente omitidos, ou então, textos cujos constituintes não estejam apresentados em sua ordenação primária;
- a verificação de aspectos mais complexos da linguagem, representados principalmente por aqueles ligados à concordância, à regência, às anáforas, aos pronomes, etc. Aqui também os autômatos adaptativos se apresentam como uma

alternativa interessante para a resolução desses problemas de dependências contextuais apresentados pelas linguagens naturais. Neste caso específico, o tratamento das dependências contextuais é feito adicionando-se ações adaptativas ao autômato utilizado para a verificação das regras de colocação.

- em um quarto estágio, são tratados problemas associados à presença de ambigüidades e não-determinismos decorrentes das inúmeras combinações legítimas permitidas pelas linguagens naturais através de permutações na ordenação de seus componentes, dando assim margem a múltiplas interpretações válidas para o texto de entrada, o que acarreta correspondentemente o aparecimento de muitos caminhos válidos simultâneos nos autômatos que implementam a análise do mesmo.

Assim, como se pode notar, é possível, com um único modelo, baseado no autômato adaptativo, efetuar todo o tratamento do texto de entrada, desde seus aspectos morfológicos mais elementares, passando pela ordenação de seus constituintes, até a verificação dos aspectos mais complexos das possíveis transposições e da interação entre as diversas partes estruturais de que é formado.

4.1 Considerações Gerais

A figura 27 representa a arquitetura de uma ferramenta que disponibiliza a um lingüista, especialista em linguagem natural, mas não necessariamente conhecedor do paradigma computacional adaptativo, uma interface através da qual possa descrever a linguagem natural através de uma gramática.

Essa ferramenta destina-se também para ser utilizada principalmente por um usuário final, seja ele homem ou máquina, interessado em obter uma floresta de árvores sintáticas,

associadas a um texto de entrada por ele fornecido, na qual cada árvore representa uma possível interpretação legítima do texto de entrada.

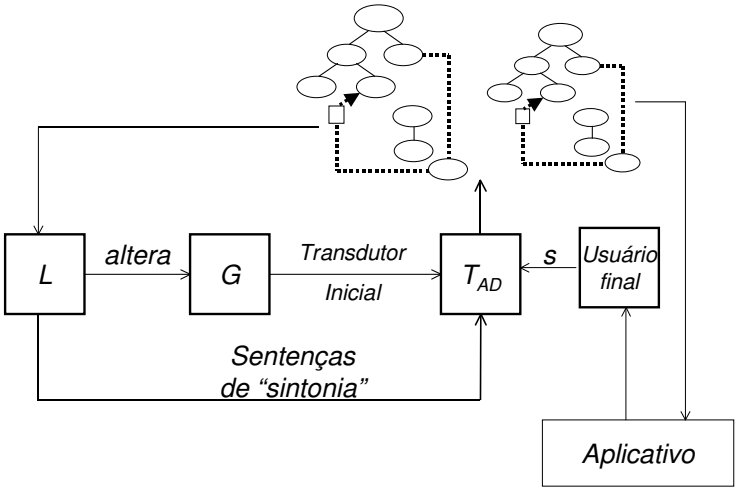

Figura 27 – Arquitetura adaptativa para processamento de linguagem natural

Nesta figura, L representa um lingüista, G uma gramática da linguagem natural e T_{AD} representa um transdutor adaptativo.

O lingüista, que é especialista em linguagem natural, tem acesso a operações através das quais pode inserir, remover ou simplesmente alterar um conjunto de regras que define a gramática G, a qual, por sua vez, descreve a linguagem L desejada.

Também é possível ao lingüista apresentar à ferramenta sentenças de "sintonia", destinadas a facilitar o ajuste da gramática, obter o conjunto das árvores de sintaxe correspondentes a essas sentenças, avaliar a gramática disponível e eventualmente efetuar modificações no seu conjunto de regras.

Em sua ***configuração inicial***, o transdutor deve ser capaz de memorizar ***um dicionário*** juntamente com as ***regras da gramática*** fornecidas pelo lingüista, estando então o transdutor

adaptativo T_{AD} em um modo de operação denominado "**treinamento**", e promovendo como resultado sua automodificação para uma nova configuração.

Isso se repete enquanto novas alterações forem sendo inseridas pelo lingüista, até que o transdutor adaptativo atinja uma ***configuração final***, na qual ele seja considerado, pelo lingüista, apto a analisar, à luz do dicionário e da gramática, textos de entrada fornecidos por um usuário final, escritos na linguagem natural a que se referem o dicionário e a gramática.

O usuário final é o elemento interessado na análise de um texto de entrada, e, do ponto de vista da ferramenta, distingue-se do lingüista porque não tem direito de efetuar alterações no conjunto de regras que definem a gramática da linguagem natural.

A seguir, os elementos da ferramenta e sua função no processamento de linguagem natural são descritos, com alguns pormenores adicionais.

4.2 Elementos Principais da Ferramenta

A ferramenta aqui proposta tem como elemento central um transdutor adaptativo, que incorpora, em sua configuração final, atingida após o devido treinamento, as funções de análise léxica e sintática, o dicionário e a gramática da linguagem, o texto de entrada a ser analisado, e as árvores de sintaxe correspondentes ao texto de entrada, de acordo com a gramática da linguagem.

Esses elementos, ligeiramente comentados a seguir, constituem as partes mais importantes da ferramenta proposta.

4.2.1 Analisadores léxico e sintático

Inicialmente identificam-se, para o transdutor adaptativo, dois grandes módulos funcionais: o analisador léxico e o analisador sintático.

O analisador léxico efetua a separação e classificação dos componentes léxicos do texto de entrada e identifica os vínculos sintáticos impostos pelos mesmos aos demais componentes da sentença.

Cabe também ao analisador léxico decompor eventualmente as palavras em seus componentes mais primitivos, separando assim os elementos constituintes das contrações, isolando os radicais das palavras, identificando seus prefixos e sufixos, e assim por diante, e dando um tratamento individualizado para cada uma dessas partículas.

É também da alçada do analisador léxico consultar o dicionário, determinando, para a palavra extraída, informações adicionais necessárias para uma eventual eliminação de ambigüidade.

Ao analisador sintático cabe primariamente o papel de reconhecedor, através do qual efetua a verificação da validade da colocação dos elementos léxicos, observada no texto de entrada.

Considerando que linguagens naturais toleram inúmeras permutações e omissões para os elementos constituintes de suas sentenças, e que tais elementos em geral podem ter estruturas similares, encontra-se sempre um problema de porte razoável quando se deseja determinar de forma rigorosa o papel de cada componente do texto analisado.

Decorre imediatamente que a convivência com essas múltiplas interpretações (ao menos, parcialmente, durante a análise) é necessária, e isso exige do reconhecedor que tenha a possibilidade de lidar com situações não-determinísticas, e do usuário, que aceite como um fato as ambigüidades da língua, recebendo eventualmente por isso mais de uma interpretação para um mesmo texto de entrada.

Superados esses problemas de não-determinismo e ambigüidade, o analisador sintático deve gerar, com base em uma descrição formal (geralmente gramatical) da sintaxe da

linguagem, e também, naturalmente, no próprio texto a ser analisado, uma ou mais árvores de sintaxe a ele correspondentes.

4.2.2 Dicionário e gramática

São bases de dados que disponibilizam à ferramenta elementos para que esta possa conduzir adequadamente o trabalho de análise da sentença em linguagem natural.

O analisador léxico se serve principalmente do dicionário para efetuar a identificação das palavras válidas da língua, a extração, a classificação morfológica e a etiquetação das palavras do texto de entrada (como resposta, pode-se esperar eventualmente mais de um resultado válido desta análise).

Já o analisador sintático se utiliza dos elementos léxicos extraídos pelo analisador léxico, bem como das regras que compõem a gramática da linguagem, para verificar a qual (eventualmente pode-se aqui também obter mais de um resultado válido) das possíveis construções sintáticas legítimas da linguagem pertence o texto em análise.

Primariamente, o dicionário pode ser visto como uma coleção de todas as palavras que a língua natural permite utilizar nas suas sentenças.

Entretanto, na prática a isso se acrescentam dados de caráter outro que não puramente morfológico, incluindo informações sintáticas e semânticas sobre a palavra em questão.

As gramáticas incluem informação sobre as seqüências que a linguagem permite utilizar na construção das sentenças.

Devem conter elementos que identifiquem os grandes componentes das sentenças, sua colocação relativa, seus elementos opcionais, a estrutura interna de cada um desses elementos e as suas regras de formação, as flexões esperadas para cada um dos componentes léxicos, etc.

Como complemento, as gramáticas devem informar acerca da interação esperada entre seus elementos, como é o caso de concordâncias e regências, omissões de elementos, etc.

4.2.3 Texto de entrada e árvores de sintaxe

Através de um arquivo de texto não-formatado, o usuário final pode alimentar o sistema com um texto de entrada, escrito em linguagem natural, esperando que em resposta o sistema construa automaticamente uma correspondente árvore de sintaxe (pode eventualmente haver mais de uma), de acordo com as regras da gramática disponível.

Naturalmente, para que haja uma correta aceitação do texto, este deve ser constituído integralmente de palavras pertencentes ao conjunto representado pelo dicionário existente, e essas palavras devem estar flexionadas de acordo com as regras de flexão impostas pela categoria lexical a que pertence, e às regras de flexão indicadas no dicionário e previstas na gramática da linguagem em questão.

Uma palavra deve ser separada de suas vizinhas por espaçadores específicos de cada linguagem, geralmente espaços em branco, sinais de pontuação, fim de linha e similares.

Cabe, obviamente, ao analisador léxico identificar os limites de cada palavra, de acordo com tais regras, e isolar essa palavra para depois iniciar sua análise.

Uma vez isolada a palavra, esta pode ser classificada e adequadamente etiquetada, para ser então empregada na construção da árvore de sintaxe do texto de entrada.

Neste ponto, a análise sintática inicia seu processo de conversão de uma seqüência de palavras classificadas e etiquetadas em uma árvore de sintaxe construída de acordo com as regras de colocação estabelecidas pela gramática da linguagem de entrada.

Árvores de sintaxe estabelecem o relacionamento estrutural entre palavras ou entre sub-árvores de sintaxe da sentença a que se referem, e são caracterizadas por apresentarem

cada qual um único nó principal, correspondente ao não-terminal raiz da gramática, nós-folha representando cada um dos terminais encontrados no texto de entrada, e nós intermediários que promovem a interligação dos demais nós.

A interligação de cada um dos nós-pai com um conjunto correspondente de nós-filhos imita a maneira como a correspondente regra de produção da gramática manda substituir um nó não-terminal (associado ao lado esquerdo da produção) por uma seqüência formada de terminais e não-terminais (relativa ao seu lado direito).

Uma árvore de sintaxe deve ter como contorno exatamente a cadeia de terminais correspondente ao próprio texto de entrada a que se refere a árvore.

Pode haver mais de uma árvore de sintaxe para uma mesma sentença da linguagem, e quando isso acontece, indica que a gramática utilizada para a definição dessa linguagem é uma gramática ambígua.

Nesse caso, árvores diferentes indicam formas diferentes (mas legítimas) de interpretação dessa mesma sentença, e a diferença entre elas está exatamente nos diferentes conjuntos de nós intermediários que as constituem, e na variedade de interconexões utilizadas na formação das diferentes árvores.

A ausência de uma árvore de sintaxe que corresponda ao texto de entrada segundo a gramática da linguagem mostra que o texto de entrada não é uma sentença corretamente formada segundo o conjunto regras gramaticais adotado para definir a linguagem natural nessa ferramenta.

4.3 Modos de operação da ferramenta

A ferramenta proposta apresenta, como foi sugerido anteriormente, dois modos de operação, a saber: o modo **treinamento** e o modo **uso**.

No modo **treinamento** a ferramenta exibe, no transdutor adaptativo que a realiza, uma configuração inicial, que permite uma primeira memorização de regras de uma gramática que caracteriza formalmente a linguagem desejada, estabelecidas por um lingüista.

Graças aos mecanismos adaptativos presentes, com base nas regras gramaticais recebidas do lingüista, o transdutor se auto-modifica em resposta a este processo de aprendizagem, de tal forma que a ferramenta resultante, nele apoiada, se vá tornando capaz de realizar cada vez mais das tarefas de análise léxica e sintática de um texto de entrada.

Sucessivas modificações nas regras vão assim sendo introduzidas através da alteração do conjunto de regras gramaticais, e as evoluções correspondentes vão sendo realizadas sobre o transdutor, até que a ferramenta se torne finalmente capaz de reconhecer e analisar um texto de entrada em todos os aspectos lingüísticos considerados essenciais da particular linguagem natural desejada, ficando assim a ferramenta pronta para ser utilizada por um não-especialista, para a análise dos seus textos específicos, em linguagem natural.

Aqui, portanto, a ferramenta está preparada para ser utilizada em modo **uso**. Naturalmente, agora se pode considerar que todas as informações importantes acerca da linguagem já são de conhecimento da ferramenta, por ter sido esta previamente treinada pelo lingüista.

O usuário final pode então fornecer à ferramenta seus textos de entrada, escritos na linguagem natural, e, com base na gramática da linguagem em questão, conhecida do sistema, pode finalmente ser produzida uma ou mais árvores de sintaxe correspondentes aquele texto.

O modo **uso**, portanto, dispensa o lingüista, e pode ser exercitado diretamente pelo usuário final, o qual faz dos resultados fornecidos pela ferramenta o uso que for mais apropriado, tais como: verificar a correção ortográfica e gramatical de seu texto de entrada; conhecer as possíveis interpretações do texto; utilizar as estruturas sintáticas possíveis, levantadas pela ferramenta, para alimentar um estágio adicional de processamento da

linguagem natural que promova a extração do conteúdo semântico do texto de entrada; alimentar um estágio de processamento encarregado de efetuar uma tradução do texto de entrada para algum outro idioma, etc.

Existe um terceiro modo de operação: trata-se do modo **expansão**. Nas linguagens naturais, são profusas as relações de subordinação. Estas dependências se dão em diferentes graus de complexidade por meio de diversos expedientes. A ocorrência de um item lexical, de uma categoria gramatical, de uma determinada flexão, etc. é gramaticalmente correta, somente se após a ocorrência do termo subordinado a este primeiro. Esta situação é facilmente tratada pelo formalismo adaptativo. Uma vez que na regra se especifique a ocorrência de um determinado símbolo é possível representar o elemento a ele subordinado. Em tempo de uso desta regra, tão logo seja detectada na cadeia de entrada o símbolo, núcleo da subordinação, pelo analisador léxico, o transdutor adaptativo se expande de forma que seja acrescentado um trecho do analisador sintático, responsável pela verificação do termos que ao núcelo se subordinam.

O mecanismo de expansão consiste em criar uma cópia do trecho da gramática que geraa o termo subordinado . A utilização do mecanismo de expansão é pois a técnica empregada para o tratamento destes não-determinismos profusos na gramática natural. É também adotado no tratamento de ambigüidades.

4.4 O tratamento de não-determinismos e ambigüidades na ferramenta proposta

Um problema recorrente no estudo de linguagens complexas, como é o caso das linguagens naturais, é a presença de situações para as quais, em resposta a um dado estímulo, o sistema nem sempre evolui obrigatoriamente para uma única situação seguinte.

Em alguns casos, é possível evitar tais situações revendo a formulação adotada para a caracterização da linguagem, e redescrevendo-a de alguma outra maneira que evite o problema.

Contudo, essa manobra raramente é possível no âmbito das linguagens naturais, e se pode afirmar com segurança que é relativamente grande a freqüência dos casos que exigem, infelizmente, que se conviva com tais situações.

Situações com múltiplas evoluções possíveis surgem todas as vezes em que for detectada, durante o tratamento da linguagem, alguma condição de ambigüidade ou de não-determinismo.

Uma das formas mais intuitivas de considerar, na prática, tais situações consiste em se estudar, em paralelo, todas as combinações alternativas por elas suscitadas, de modo que, se alguma dessas combinações for válida, tal combinação leve à construção de uma árvore de derivação completa e correta.

Equivalentemente, é possível obter o mesmo resultado fazendo-se um a um, seqüencialmente, o estudo individual de apenas uma dessas possibilidades de cada vez, como se as demais não existissem, coletando ao final de cada um desses estudos, a eventual árvore completa que este caso particular origina.

Como no caso paralelo, caso se adote essa técnica, o processamento sucessivo das diversas combinações válidas de alternativas deve ser feito até que todas elas tenham sido verificadas, e todas as árvores válidas, coletadas.

Naturalmente, nessa opção seqüencial, os casos de insucesso na construção da árvore leva a descartar aquela combinação de possibilidades, e não à rejeição da entrada.

A rejeição do texto de entrada só deverá acontecer, neste caso, se todas as combinações consideradas inicialmente como válidas falharem em produzir árvores corretas.

Do ponto de vista de implementação, é possível considerar caso a caso tais condições indesejáveis de não-determinismo ou ambigüidade, ou então, criar um mecanismo geral para efetuar seu tratamento de maneira mais automática.

Neste último caso, disponibiliza-se um procedimento geral de tratamento de tais fenômenos, o qual pode então ser ativado sempre que necessário, sem que haja necessidade que se faça individualmente um detalhamento específico dos pormenores de cada caso particular.

Em (NETO; MORAES, 2002), é apresentada uma proposta adaptativa para o tratamento de não-determinismos e ambigüidades onde se busca o conjunto de todas as interpretações possíveis para uma dada sentença.

Para isso, o transdutor adaptativo deve ser construído de tal forma que as construções normais sejam aceitas da forma usual, e que as situações de não-determinismo ou de ambigüidade, uma vez identificadas, criem, para serem executadas em paralelo, cópias da gramática, cada uma das quais capaz de reconhecer uma das possíveis construções sintáticas válidas.

O paralelismo da operação desses autômatos é simulado através da sua execução alternada, passo a passo, em que cada um deles efetua uma transição por vez, correspondente ao consumo de um símbolo da cadeia de entrada.

As diversas versões da gramática vão sendo executadas simultaneamente, sendo descartadas aquelas que não permitirem o prosseguimento do reconhecimento, e sendo consideradas vitoriosas aquelas que conseguirem esgotar a cadeia de entrada em uma situação final.

Havendo mais de uma máquina vitoriosa, para a mesma sentença, isso será indicativo de que a sentença em questão tem mais de uma interpretação válida, decorrente da ambigüidade presente na linguagem.

Uma das premissas fundamentais em (NETO; MORAES; 2002) é pois que o tratamento adequado das situações de não-determinismos e ambigüidades faça uso de novas instâncias de uma sub-gramática original. Essa situação corresponde nas linguagens de programação ao tratamento de macros. A proposta adaptativa para o tratamento de macros é apresentada na seção 4.4.8 em (NETO, , 1993).

Suscitada pela busca de um mecanismo adaptativo de cópia, necessário ao tratamento de não-determinismos e ambigüidades efetuou-se uma análise dos referidos mecanismos. Concluiu-se que as mesma técnicas podem ser empregadas em processamento de Linguagem Natural. Tal análise é relatada no capítulo 4 da presente tese.

4.5 Descrição dos Componentes da Ferramenta

A figura 28 ilustra, em um primeiro nível de detalhamento, os blocos funcionais da ferramenta destinados ao tratamento sintático da linguagem natural, a saber: o analisador léxico, o analisador sintático e o módulo que efetua o tratamento das ambigüidades e não-determinismos detectados pela análise léxica ou pela análise sintática.

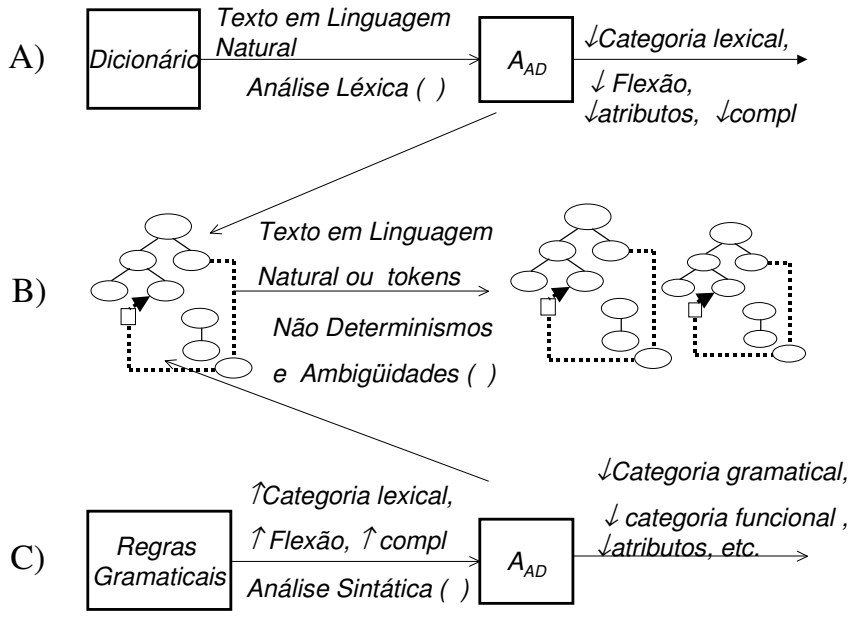

Figura 28 – Blocos Funcionais: A) Análise Léxica B) Tratamento de Não-Determinismos e Ambigüidades C) Análise Sintática

Observe-se que, neste nível superior da abstração, nada é imposto acerca da natureza da linguagem de entrada da ferramenta, e, portanto, o raciocínio aqui desenvolvido, embora seja melhor aplicado a linguagens de alta complexidade, pode, na realidade, ser aplicado a qualquer tipo de linguagem, mesmo que não seja uma linguagem natural.

Em modo **treinamento**, os blocos funcionais podem ser empregados para tratarem a meta-linguagem através da qual a ferramenta recebe do lingüista as informações sobre a linguagem natural a ser processada, e essa meta-linguagem, cuja estrutura é livre de contexto, é com muita certeza muito mais simples que qualquer linguagem natural que através dela costuma ser definida.

A descrição, a seguir, dos componentes representados pelos blocos funcionais da ferramenta, referem-se mais propriamente ao seu funcionamento em modo **uso**.

4.5.1 O analisador léxico

Em 28.A) representa-se o ***analisador léxico***, bloco funcional da ferramenta que é responsável pela extração das palavras do texto fornecido pelo usuário final, pela sua classificação segundo a categoria lexical a que pertence, e sua decomposição em partes menores, correspondentes aos desmembramentos de contrações e de flexões, prefixos e sufixos incorporados nas palavras extraídas.

Para reconhecer e classificar as palavras, o analisador léxico faz uso de um dicionário cujo conteúdo deve conter informações que permitam determinar todas as categorias lexicais possíveis para cada palavra.

O resultado da execução do analisador léxico compõe-se de um ou mais tokens, relativos à primeira das palavras ainda não extraídas do texto, devidamente separada do texto de entrada, desmembrada em seus componentes primários, apropriadamente classificada (categoria lexical, informação sobre o tipo de complementos exigidos) e acompanhada dos seus atributos de flexão (gênero, número, grau, pessoa, tempo, modo, etc.), e de informações relativas ao quadro de subcategorização do item lexical.

É conveniente notar que no dicionário diversas informações adicionais, de cunho não puramente morfológico, costumam ser obtidas sobre a palavra extraída, tais como as exigências contextuais da palavra, em matéria de concordância e de regência, informações sobre seus possíveis significados em diversos contextos, informações sobre os seus usos mais freqüentes, expressões idiomáticas nas quais costuma figurar, etc.

É importante notar a forte interação necessária que deve existir entre o analisador léxico e o dicionário, por um lado – é no dicionário que o analisador léxico encontra a maior parte das informações lingüísticas sobre a palavra analisada, e entre o analisador léxico e o

analisador sintático, por outro – o analisador sintático é um forte usuário das informações obtidas na análise léxica.

4.5.2 O analisador sintático

Em 28.C) representa-se o analisador sintático, bloco funcional da ferramenta que se responsabiliza pela análise da estrutura do texto de entrada, e pela construção de todas as possíveis árvores de sintaxe, que possam ser geradas pela gramática da língua natural em questão a partir das saídas do bloco funcional que executa a análise léxica da sentença fornecida como texto de entrada pelo usuário final.

Há na literatura muitos métodos relatados para a realização da análise sintática para uma linguagem descrita por uma gramática. (AHO, ULLMAN, 1972); (ALLEN, J., 1995). Em (JURAFSKY; MARTIN, 2000) e (MATTHEWS, 1998.) são citados inúmeros trabalhos.

Entre as estratégias de análise sintática mais utilizadas, destacam-se: a análise descendente, a análise ascendente e a análise com o auxílio de reconhecedores ou autômatos.

Em qualquer das análises, tem-se uma área de trabalho, na qual se vai construindo a árvore à medida que isso for possível a partir da configuração corrente da árvore, do conteúdo do texto de entrada e de alguma regra gramatical aplicável no momento.

Distinguem-se as análises descendente e ascendente quanto ao não-terminal específico que, com base nas regras gramaticais aplicáveis, é escolhido para ser substituído, na área de trabalho, num determinado momento da análise.

Assim, ***analisadores descendentes*** escolhem sempre o não-terminal mais à esquerda ainda não substituído para sobre ele aplicar uma regra gramatical em que ele conste como membro esquerdo, substituindo-o na área de trabalho, pelo lado direito da regra em questão.

Isso faz com que a árvore de sintaxe seja construída partindo sempre da raiz, em direção às folhas, e que o ponto em que é feita a substituição seja um ponto de limite entre a parte esquerda da árvore, cujas folhas são todas constituídas de terminais, e a parte direita da mesma, cujo contorno não é formado apenas de terminais, mas portanto apresenta ainda nós não-terminais a serem substituídos.

Para dar partida à operação de um analisador descendente, inicia-se a área de trabalho apenas com o não-terminal que corresponde à raiz da gramática da linguagem desejada, correspondendo a iniciar o trabalho apenas com a raiz da árvore de sintaxe.

No prosseguimento da análise descendente, escolhe-se para ser substituído sempre o nó não-terminal cuja posição na árvore seja aquela mais à esquerda no contorno da árvore.

Em outras palavras, visto pela perspectiva da forma sentencial que constitui o conteúdo da área de trabalho em cada momento, procura-se substituir o não-terminal mais à esquerda dessa forma sentencial.

A escolha da substituição a ser aplicada consiste em consultar a gramática, em busca de uma regra de substituição para tal não-terminal, mas que seja compatível com o próximo token extraído do texto de entrada pelo analisador léxico.

Por outro lado, ***analisadores ascendentes*** escolhem primeiramente para ser substituído o não-terminal mais à direita da forma sentencial contida na sua área de trabalho.

Inicialmente, preenche-se a área de trabalho com o texto de entrada devidamente convertido pelo analisador léxico à forma de uma seqüência de tokens.

Um dispositivo tomador de decisão, específico para cada gramática e linguagem, e construído (usando-se técnicas clássicas, (cuja discussão extrapola o escopo deste trabalho) geralmente com base em um autômato finito, varre a área de trabalho, em busca de reduções a aplicar sobre a forma sentencial corrente.

Uma *redução* consiste na substituição, na forma sentencial, de uma ocorrência mais à direita de algum padrão correspondente ao lado direito de alguma regra gramatical da gramática, pelo não-terminal associado ao lado esquerdo da regra que estiver sendo aplicada.

Após cada redução, repete-se o processo, procurando-se aplicar sucessivas reduções sobre a forma sentencial contida na área de trabalho, até que esta seja reduzida, finalmente, a um único não-terminal, correspondente à raiz da gramática.

A coleta das reduções efetuadas permite a construção da árvore de sintaxe do texto analisado pela aplicação da seqüência coletada de todas as decisões tomadas pelo dispositivo tomador de decisões do analisador sintático ascendente.

Uma terceira possibilidade, que foi adotada para a ferramenta aqui proposta, consiste na utilização de um reconhecedor como elemento básico do analisador.

Pelo fato de que em sua operação um reconhecedor se limita a aceitar ou não suas cadeias de entrada como como sendo parte da linguagem, torna-se conveniente enriquecê-lo com algumas funções adicionais, em particular, com a capacidade de também gerar saídas, o que caracteriza o dispositivo assim construído como sendo um transdutor.

Para a ferramenta proposta neste trabalho, escolheu-se como técnica a aplicar a utilização de um transdutor baseado em autômatos de pilha estruturados adaptativos como elemento primário responsável pelo reconhecimento sintático da linguagem.

Devido à natureza complexa das linguagens naturais (ambigüidades, recursividade à esquerda), nem todos os tipos de autômato se servem para essa função, razão pela qual se optou pelo emprego de autômatos de pilha estruturados adaptativos.

Estes dispositivos (NETO, 1993), apresentam poder de máquina de Turing, e têm a desejável propriedade de permitir o tratamento simplificado de linguagens regulares ou livres de contexto apenas utilizando suas transições básicas e sua pilha sintática.

Por outro lado, caso sejam necessários maiores recursos do autômato, por causa da complexidade da linguagem, os autômatos adaptativos permitem também representar o tratamento de dependências contextuais, necessárias ao processamento de linguagens dependentes de contexto, como é o caso das linguagens naturais.

Para que isso se torne possível, esses dispositivos fazem uso de ações adaptativas, destinadas a proporcionar ao autômato a capacidade de se auto-modificar, podendo eles, através desse recurso, memorizar informações e alterar seu próprio comportamento, adaptando-o às necessidades de cada particular situação.

Disso resulta um dispositivo que, a partir de uma configuração inicial genérica, em que qualquer texto de entrada poderia, potencialmente, ser aceito, tenha a capacidade de modificar-se à medida que elementos específicos do texto de entrada vão sendo encontrados, adaptando-se assim às suas particulares necessidades.

Através da incorporação ao autômato das informações adequadas, extraídas dos elementos do texto, à medida que a análise desse texto de entrada vai se desenvolvendo, o autômato vai se especializando correspondentemente.

Dessa maneira, torna-se possível projetar as ações adaptativas de auto-modificação do autômato concebidas de tal forma que opções lingüísticas não utilizadas possam ir sendo excluídas das atenções do autômato, enquanto aquelas que estejam em uso sejam enriquecidas com as opções viáveis apenas.

O resultado do uso dessa estratégia é que, dinamicamente, o autômato adaptativo assim construído permanece sempre em sintonia com o particular texto de entrada que está sendo analisado, e que partes do autômato que não seriam percorridos nem sequer são mantidas no dispositivo, evitando assim o gasto desnecessário de áreas de armazenamento.

4.5.3 O tratamento de ambigüidades e não-determinismos

Na figura 28.C) representou-se o módulo que efetua o tratamento das ambigüidades e não-determinismos.

Embora não se trate de uma componente cuja necessidade seja evidente num primeiro momento, é de suma importância a sua presença no sistema, pois permite uniformizar e, até um certo ponto, automatizar o tratamento de todas as situações em que mais de uma decisão válida possa ser tomada a partir de um mesmo estímulo, em algum ponto específico da análise (léxica ou sintática).

Ao ser acionado, este módulo cria as instâncias que forem necessárias, da análise em andamento, para se levar em consideração, separadamente, cada um dos casos possíveis detectados.

No caso da análise léxica, este componente de tratamento de ambigüidades e não-determinismos deve ser acionado sempre que o token extraído apresentar mais de uma classificação possível.

No caso da análise sintática, o tratamento de ambigüidades e não-determinismos deve ser acionado sempre que mais de uma regra gramatical puder ser aplicada, ou, equivalentemente, quando, a partir do estado corrente do autômato, mais de um caminho legítimo for encontrado para o token em uso.

Convém lembrar que o uso automático desses procedimentos de tratamento de ambigüidades e não-determinismos só se torna possível se seu acionamento for implicitado no funcionamento geral dos dispositivos de análise.

Assim, o transdutor que implementa as análises léxica e sintática subentende a ativação automática do analisador léxico todas as vezes que um token é consumido por

alguma transição do autômato subjacente do transdutor sintático, de tal modo que sempre se tenha disponível um próximo token a ser analisado.

Da mesma maneira, deve-se subentender a ativação dos mecanismos para o tratamento adequado dos não-determinismos ou das ambigüidades correspondentes às situações em que o analisador léxico ou o analisador sintático se depararem com múltiplas decisões válidas, igualmente aplicáveis no momento.

Isto é o que acontece todas as vezes que uma resposta múltipla for gerada pelo analisador léxico (mais de uma classificação possível para uma mesma palavra) ou então pela consulta à gramática, em busca das regras aplicáveis (mais de uma regra, todas igualmente aplicáveis), ou ainda pela detecção de uma situação não-determinística no autômato subjacente ao transdutor utilizado).

Em qualquer desses casos, cada uma das possibilidades deve ser analisada, qualquer que seja a técnica adotada: uso de paralelismo ou da serialização das situações, ou o uso de threads na implementação, permitindo explorar algum paralelismo disponível na plataforma existente ou então promovendo a simulação de tal paralelismo. Como já citado, em (NETO; MORAES, 2002) adotou-se a simulação do paralelismo

4.6 As bases de dados

Para ser possível à ferramenta efetuar adequadamente a análise de uma linguagem, uma descrição completa de todos os aspectos desta linguagem que sejam relevantes para a análise desejada deve estar formalmente representada no sistema, disponível em alguma notação por ele conhecida.

Tais descrições podem ser apresentadas de inúmeras maneiras alternativas, como, por exemplo, entre muitas outras possibilidades, na forma de functores, na forma de expressões

regulares, na notação de Wirth, na forma de gramáticas adaptativas, na de conjuntos de produções livres ou dependentes de contexto, na forma de autômatos finitos, ou de pilha, ou adaptativos, etc.

Todas essas formas são válidas, podem ser indiferentemente utilizadas, e se equivalem mutuamente, guardados os seus níveis relativos de expressividade, e a escolha dentre elas pode ser feita com base em algum critério arbitrário, usualmente baseado nas disponibilidades e nas conveniências do projeto.

A equivalência teórica dos diversos formalismos computacionais anteriormente citados, permite converter, direta ou indiretamente, cada um deles em uma versão equivalente, denotada no formato dos demais formalismos, entre os quais, os autômatos de pilha estruturados adaptativos.

Em (TANIWAKI, 2000) encontram-se alguns resultados obtidos nesse sentido, que mostram algumas das equivalências e constatam a viabilidade do uso de técnicas adaptativas no processamento de linguagens naturais.

Outro esforço neste sentido, no qual se demonstrou viável a utilização de tecnologia adaptativa na construção de analisadores léxicos ou etiquetadores automáticos para a língua portuguesa, encontra-se em (MENEZES, 2000)

O procedimento da análise de linguagem natural aqui utilizado, empregando os transdutores adaptativos, inspira-se nos métodos inicialmente apresentados em (NETO, 1993) para a resolução de dependências de contexto em linguagens de programação, detalhado no capítulo anterior.

Ainda, convém, naturalmente, destacar que uma ferramenta como a que está sendo proposta disponibiliza a seus usuários uma interface em linguagem natural que pode ser utilizada por outros aplicativos, tal como foi apresentado esquematicamente na figura 27.

Com a ajuda deste tipo de recurso, o usuário final dos aplicativos pode operá-los fornecendo-lhes sentenças em linguagem natural em lugar de instruções especiais, expressas em alguma linguagem de comandos especialmente desenvolvida para o aplicativo.

Por sua vez, para o lingüista, que manipula os aspectos ligados à formalização da linguagem natural utilizada, a ferramenta deve dar acesso a alguma meta-linguagem, tipicamente alguma notação gramatical, empregada para expressar a estrutura da linguagem.

Nessas meta-linguagens, regras de construção de sentenças permitem desenvolver uma descrição rigorosa da estrutura sintática da linguagem, da qual é possível determinar com precisão o comportamento esperado dos seus analisadores léxico e sintático.

Adicionalmente, a existência, em tais gramáticas, de informações ligadas à forma de identificação de dependências contextuais, no texto de entrada, permite escolher a forma mais adequada para sua detecção em um particular texto de entrada, e subseqüente tratamento.

Para a ferramenta proposta, são identificados os seguintes repositórios principais de informação: o dicionário, a gramática ou autômato, o texto de entrada e a árvore de sintaxe.

Nesta proposta, embora tais elementos estejam apresentados conceitualmente de forma clássica, fazem extensivo uso da capacidade de auto-modificação característica dos dispositivos adaptativos, e se apresentam portanto com uma conotação dinâmica.

Discorre-se em seguida sobre os elementos em questão.

Dicionário:

No *dicionário* estão coletadas e organizadas todas as palavras da linguagem natural com as quais se pode construir sentenças. Muitos dicionários não colecionam todas as possíveis palavras, mas apenas as chamadas "formas de dicionário", excluindo assim um número muito grande de repetições devidas às inúmeras possíveis flexões que as palavras

podem sofrer ao serem empregadas em sentenças da língua, bem como os principais atributos (etiquetas) associados, para uso da análise léxica.

No dicionário figuram também as categorias gramaticais que a palavra costuma assumir nas sentenças que a empregam, bem como os principais atributos (etiquetas) associados, para uso da análise léxica.

Dessa maneira, o analisador léxico consulta o dicionário em busca de uma palavra, e uma vez localizada essa palavra, extrai as correspondentes informações sobre as categorias gramaticais a que pode pertencer, suas possíveis flexões em cada caso, bem como sobre suas exigências léxicas e sintáticas.

Gramáticas e autômatos:

Para descrever formalmente a linguagem, pode-se escolher um mecanismo formal gramatical ou então um dispositivo reconhecedor (autômato).

A teoria garante que tais formalismos têm o mesmo poder descritivo, e portanto são equivalentes, logo usar um ou outro é apenas uma questão de conveniência.

Naturalmente, o uso simultâneo das duas formas de descrição da linguagem não é conveniente por causa das redundâncias inevitáveis, além das dificuldades de manutenção que essa prática certamente acarreta, no caso de alguma alteração se mostrar necessária.

Caso se opte por uma gramática, podem-se dividir as regras gramaticais em dois níveis (em geral na prática não são apresentados em separado, mas formam um único conjunto de regras): um nível superior, que se ocupa das grandes estruturas da língua, e um outro nível, inferior, que descreve os detalhes das construções sintáticas permitidas em cada caso.

Caso seja escolhido um reconhecedor (autômato), considerando que a dificuldade de construir diretamente um autômato é bem maior que o de construir uma gramática, os lingüistas em geral preferem descrever linguagens usando formulações gramaticais, e portanto

na prática costuma-se partir de uma gramática, fornecida pelo lingüista, convertendo-a se necessário em autômato, usando para isso algum algoritmo disponível.

Se isso for verdade, torna-se possível automatizar esse processo, incluindo-se na ferramenta uma implementação de tal algoritmo, e fornecendo à ferramenta, em seu modo de **treinamento**, a gramática desejada, para ser convertida em autômato.

Posteriormente, um autômato, equivalente à gramática, construído e instalado pela própria ferramenta enquanto operava em modo **treinamento**, irá comandar, em modo **uso**, a operação dos procedimentos de análise do texto de entrada de acordo com o estabelecido na gramática originalmente fornecida à ferramenta.

Texto de entrada:

O processamento de textos em linguagem natural exige, obviamente, que lhe seja fornecido algum texto escrito em linguagem natural, para ser processado.

Esse texto deve ser apresentado à ferramenta pelo usuário final, sempre em modo **uso**, e o único módulo que dele extrai informações é o analisador léxico.

Tais informações são codificadas pelo analisador léxico para uso subseqüente na análise sintática.

O analisador sintático, por sua vez, pode efetuar outras transformações sobre esse conjunto de informações, tendo em vista a construção da árvore sintática da sentença em análise.

Na codificação feita pelo analisador léxico, as palavras são convertidas em categorias lexicais e estas, associadas a outros atributos extraídos da própria palavra e de informações complementares, obtidas nos dicionários.

Árvore de sintaxe:

Uma das metas mais importantes que se deseja alcançar quando se constrói um programa para o processamento automático de linguagem natural é a obtenção de uma árvore de sintaxe para cada texto de entrada fornecido.

À luz de uma gramática, pode-se visualizar a árvore sintática como uma árvore de derivação do texto de entrada, derivação essa feita com base na gramática.

Assim, a raiz da gramática será o nó principal da árvore de derivação, e as folhas dessa árvore corresponderão às palavras extraídas do texto de entrada.

Os nós intermediários correspondem a não-terminais da gramática, que tenham sido utilizados para a derivação da sentença.

Esses não-terminais intermediários são escolhidos sempre que uma regra gramatical for aplicada: o lado esquerdo da regra determina o não-terminal, e o seu lado direito, a substituição indicada pela regra.

Na árvore, o nó correspondente ao não-terminal presente no lado esquerdo da regra terá como nós-filhos os elementos contidos no lado direito da regra.

4.7 Representação gramatical da linguagem natural

Adota-se neste trabalho, para a representação gramatical da linguagem natural, uma notação similar à própria declaração de uma Função Adaptativa.

No caso particular aqui considerado, o problema a resolver consiste em determinar a estrutura sintática de um texto de entrada fornecido, e as premissas que devem ser levadas em conta nesta determinação têm como base os fatos contidos no dicionário, as palavras extraídas na forma de tokens pelo analisador léxico, e as maneiras legítimas de encadeamento, permitidas para esses elementos.

As regras que definem a linguagem prestam-se primordialmente a descrever as formas permitidas para a construção de sentenças válidas, de acordo com as estruturas que caracterizam a linguagem que interessa, em particular, ao usuário da ferramenta.

A raiz da gramática deve ser especificada da seguinte forma:

identificador.

Exemplo:

periodo.

As regras, que definem uma estrutura sintática devem ser especificadas da seguinte forma:

identificador_regra (lista opcional de parâmetros) =

{ lista_identificadores $\rightarrow$ { lista_regras }

O identificador_regra e o escopo onde a mesma é armazenada identificam juntos uma determinada regra.

Os parâmetros destinam-se a representar os atributos de uma determinada categoria, tais como gênero, número, grau, pessoa, papel temático, etc. Seus identificadores são precedidos pelo símbolo "&".

Os identificadores representam as categorias lexicais e gramaticais. Assim, quando se deseja representar uma estrutura sintática, sua especificação pode ser indicada como no exemplo seguinte.

periodo ( ) = { ip ponto_final }

periodo ( ) = { ip α_coord $\rightarrow$ { periodo ( ) } ponto_final }

Estas regras especificam um período simples e um período composto por coordenação, respectivamente.

Os parâmetros podem ser empregados quando se deseja especificar, por exemplo, atributos dos constituintes da oração ou mesmo dos itens lexicais. Seguem-se alguns exemplos:

dp (&gen, &num) = { det &gen &num → {np(&gen, &num)} }

Esta regra especifica o grupo determinante, com a presença obrigatória do determinante, o qual subcategoriza uma instância do grupo nominal, o qual por sua vez concorda em gênero e número com o determinante.

np(&gen, &num) = {adj &gen &num coord → {np(&gen, &num)}}

np(&gen, &num) = { adj &gen &num subst &gen &num}

np(&gen, &num) = {subst &gen &num}

Estas regras especificam o grupo nominal, com a presença de adjetivos que podem ocorrer de forma coordenada, ou um adjetivo justaposto a um substantivo, ou ainda um único substantivo.

Representação da componente léxica da linguagem.

A representação da componente léxica pode fazer uso da mesma sintaxe da metalinguagem empregada para a componente sintática da linguagem. Naturalmente depende apenas dos fenômenos lingüisticos que se deseja representar. O exemplo que se segue, apresenta uma forma simplificada de se representar um determinante. No capítulo seguinte apresenta-se a representação do fenômeno de subcategorização na representação da componente lexical

o ( ) = {det masc sing }

o ( ) = {pron pessoal acusativo masc sing }

o = {pron demonstrativo masc sing}

4.8 Primeiro detalhamento do nível de implementação

A "implementação" da ferramenta capaz de realizar as tarefas de análise léxica e sintática em linguagem natural é efetuada automaticamente pelo transdutor adaptativo em sua ***configuração inicial***, quando no sistema são inseridos o dicionário e o conjunto de regras gramaticais, ou seja, no modo treinamento.

Estas bases de dados são fornecidas através de um texto de entrada e são interpretadas pela ferramenta como um conjunto de identificadores (terminais e não-terminais da gramática da linguagem natural) e meta-símbolos. Tais identificadores e meta-símbolos são armazenados e associados a ações adaptativas adequadas para que possam ser executadas no modo uso da gramática.

A alternância entre o modo de treinamento e o modo uso se faz de forma muito simples no formalismo adaptativo. Ainda, faz parte do formalismo adaptativo, ações de inserção, remoção e consulta a regras. Dessa forma, é possível alternar o modo de operação treinamento e uso sempre que o Lingüista assim achar conveniente e ao mesmo será possível inserir ou remover as regras existentes.

Seguem-se detalhes de implementação para análise sintática e léxica

4.8.1 Implementação da análise sintática

Em geral, a análise sintática se baseia na composição da informação fornecida pelo analisador léxico, e proveniente do texto de entrada, com a informação contida nas regras que definem a gramática da linguagem natural.

Na presente proposta, a gramática original da linguagem, tal como inserida pelo Lingüista, é associada a um conjunto de ações adaptativas de forma que:

- no modo "uso" da gramática, a cada acesso a uma regra seja criada uma instância da mesma. Na arquitetura que se propõe, quando o analisador sintático tiver de acessar uma nova regra gramatical, esta não é meramente empregada para a verificação da estrutura sintática da cadeia de entrada. Na verdade, em cada acesso a uma regra gramatical, é criada uma nova instância da mesma, a qual efetivamente se prestará à execução da análise sintática. A criação de novas instâncias de regras gramaticais, já é uma tarefa que por si colabora no tratamento eficiente dos não-determinismos e ambigüidades sintáticas.
- seja criada uma lista de ponteiros que associa os lados esquerdo e direito da produção; Esta lista de ponteiros, orienta o processo de derivação da análise sintática e portanto representa um trecho da árvore; Considerando-se todas as regras fornecidas por um Lingüista, tem-se um grafo conectado representando todas as possibilidades disponíveis. No modo uso, a sentença de entrada selecionará trechos desse grafo, criando cópias do mesmo e construindo dessa forma a árvore.
- As ações adaptativas devem promover a comutação automática entre o processo de derivação e a leitura da cadeia de entrada. Isto significa que se tarefa de derivação é bem conduzida, inserirá um símbolo não-terminal na cadeia de entrada e por outro lado, a leitura do mesmo ativará o prosseguimento da derivação da sentença.
- ambigüidades e não-determinismos que se manifestam na cadeia de entrada sejam detectados e tratados convenientemente. Para isto, na leitura da cadeia de entrada, o primeiro símbolo presente na cadeia de entrada é detectado e interpretado como o delimitador de uma nova construção sintática. Na próxima ocorrência do mesmo, que é inicializada pelo mesmo símbolo. As ocorrências dos símbolos presentes entre a

ocorrência destas duas instâncias de símbolos, são copiados em uma esturutra auxiliar e reconduzidos ao analisador léxico (da configuração inical).

Na análise sintática, no modo uso, a ocorrência de um identificador na forma não marcada, implica que a regra por ele representada deverá promover a ampliação da gramática, seguida da leitura da cadeia de entrada. Ao final da leitura da cadeia de entrada, o símbolo na forma marcada é inserido na cadeia de entrada.

A alternância entre o acesso ao texto de entrada e a criação de uma instância de uma regra gramatical, configura uma operação que se faz ora de forma redutiva, ora de forma descendente.

As produções fornecidas pelo Lingüista no modo treinamento são convenientemente transformadas em Autômatos de Pilha Adaptativos. Mais precisamente, assumindo-se que o lado esquerdo da produção seja constituída de um só não-terminal correspondente ao nó pai em uma árvore, a este é associado uma lista de ponteiros correspondentes aos nós filhos. Assim, em tempo de uso da regra, esta lista de ponteiros poderá efetivamente orientar a construção da árvore.

4.8.1.1 OPERAÇÃO EM MODO "TREINAMENTO" :

Em geral, a análise sintática se baseia na composição da informação fornecida pelo analisador léxico, e proveniente do texto de entrada, com a informação contida nas regras que definem a gramática da linguagem natural.

4.8.1.2 OPERAÇÃO EM MODO "USO":

Tendo a ferramenta sido previamente treinada com as informações léxico-sintáticas da linguagem, o analisador sintático deverá estar pronto para analisar os textos de entrada que lhe forem submetidos, pois nesta ocasião o dicionário e a gramática já serão de seu conhecimento.

É aqui que todas as partes da ferramenta colaboram para a análise do texto de entrada, conforme esboçado a seguir.

Toda vez que for solicitado, o analisador léxico extrairá e classificará a próxima palavra do texto de entrada, disponibilizando o conjunto de todas as possíveis interpretações morfológicas para a palavra, acompanhadas dos atributos de flexão e das exigências morfo-sintáticas associadas a cada interpretação morfológica.

Toda vez que receber do analisador léxico informações sobre a próxima palavra do texto de entrada, o analisador sintático irá pesquisar, para cada uma de suas possíveis interpretações, a(s) possível(is) regra(s) gramatical(is) ou, correspondentemente, a(s) possível(is) transição(ões) do autômato reconhecedor à(s) qual(is) essa nova palavra se ajuste, respeitado o histórico de análise das palavras já analisadas pela ferramenta.

Toda vez que o analisador léxico ou o analisador sintático se deparar com mais de um caminho a seguir, que seja legítimo para a cadeia de entrada correntemente em análise, os mecanismos de tratamento de não-determinismos e ambigüidades deve ser ativado para que todos os casos possíveis sejam independentemente analisados.

Toda vez que o mecanismo de tratamento de não-determinismos e ambigüidades for ativado, deve receber do elemento que o ativou informação suficiente sobre as diversas opções a serem simultaneamente consideradas, de modo tal que seja promovida a instanciação de todas elas para o prosseguimento da análise: caso uma (ou mais) das instâncias consiga analisar integralmente o texto de entrada, a(s) árvore(s) sintática(s) correspondente(s) é(são) considerada(s) válida(s) para aquele texto de entrada. Se, em qualquer momento, alguma delas

não conseguir prosseguir na análise, então deverá ser desconsiderada, visto que se trata de um texto de entrada que não é aderente à gramática adotada.

Apresenta-se a seguir um método de análise para cada um dos elementos do software de processamento de linguagem natural.

4.8.2 Implementação da análise léxica

O Analisador Léxico, é implementado como uma parte do transdutor adaptativo, que é ativado sempre que o próximo símbolo na cadeia de entrada for uma palavra do dicionário, e não um símbolo não-terminal.

As palavras são buscadas e classificadas segundo na seqüência e conforme as regras fornecidas pelo Lingüista.

Em outras palavras, a especificação da arquitetura do etiquetador ou analisador morfológico. é determinada pelas regras fornecidas pelo especialista em Linguagem Natural.

Uma possibilidade para esse etiquetador é de que ele venha a fornecer o conjunto de todas as possíveis classificações para uma dada palavra, e em diversas situações, estabelecer, para cada caso, as possíveis exigências que a palavra impõe acerca de concordâncias ou de regências (sub-categorização). Exemplo

De qualquer maneira, a cada entrada pode-se obter um conjunto vazio, unitário ou múltiplo de saídas. Aqui o analisador léxico pode interagir com o mecanismo de tratamento de ambigüidades e não-determinismos, para que eventuais múltiplas interpretações possam ser tratadas

4.8.3 Comunicação entre a analise léxica e a sintática:

Determinado o conjunto possível de classes para uma palavra, o analisador léxico substitui a palavra recém analisada pela classe correspondente, acrescida das etiquetas correspondentes, deixando essa informação disponível para uso do analisador sintático.

Uma possivel opção é a de utilizar para isso a própria cadeia de entrada (que agora se transformou em uma cadeia de trabalho, ou de rascunho: considerando que duas ou mais diferentes interpretações léxicas de uma mesma palavra devem, cada uma de sua vez, provocar diferentes alterações sobre essa cadeia de entrada, e que portanto tais alterações devem ser mutuamente exclusivas, então na ocasião em que uma das opções estiver sendo adotada, as demais devem ser ignoradas, como se não existissem). Isto é possível graças aos mecanismos de delimitação de escopos indicadas no capítulo 2.

De qualquer modo, mais uma vez, para evitar complexidades desnecessárias no método empregado, é preciso escolher, com base em algum critério sólido, uma opção apenas de cada vez, e desconsiderar definitivamente todas as demais possibilidades.

4.8.4 Implementação dos mecanismos de dependências (sintáticas) de contexto:

Para que todas essas operações possam ser realizadas de forma interativa, e para que os mecanismos de análise possam ser executados adequadamente, modificando dinamicamente o formalismo (autômato ou gramática adaptativa) escolhido para a representação da linguagem, uma série de ações adaptativas devem existir que garantam a modificação apropriada do formalismo subjacente, de acordo com as necessidades determinadas pelos mecanismos de reconhecimento e análise adotados.

Assim, as diversas operações adaptativas que foram descritas nos capítulos anteriores preenchem exatamente essa função, e cada uma delas, na sua ocasião mais propícia, irá efetuar as alterações no formalismo subjacente de forma que as informações, de caráter sintático ou então referentes às dependências de contexto, que estiverem sendo tratadas possam ser devidamente incorporadas ao formalismo que implementa a análise da linguagem, na forma de ampliações ou de alterações da gramática ou do autômato correspondente.

4.9 Considerações Finais

Este capítulo apresentou diversas considerações complementares no que diz respeito a uma arquitetura adaptativa.

A proposta desta tese é que a sua implementação é aquela apresentada em (NETO, 1993), de forma que se possa auferir o desempenho diferenciado do tratamento de dependências de contexto advindos do formalismo adaptativo, conforme minuciosamente identificado nos capítulos anteriores.

Assim, na configuração incial do transdutor adaptativo, existe um analisador léxico, constituído de:

a) transdutores responsáveis pelo reconhecimento de palavras reservadas. Nesta tese se propõe que tais palavras reservadas sejam os atributos gramaticais tais como gênero (masculino, feminino, neutro), pessoa (1, 2, 3), papel temático, etc.
b) um transdutor léxico, onde figuram transições adaptativas que consomem símbolos isolados, tais como •, ⊥, ■ , &, ou ainda, transições que consomem um único símbolo ASCII com inserção do respectivo token.
c) um transdutor capaz de tratar identificadores de regras
d) um transdutor capaz de tratar identificadores de parâmetros

Ainda, o sistema deve apresentar um transdutor sintático capaz de verificar a sintaxe da metalinguagem utilizada para a representação da Linguagem Natural, memorizar e criar cópias das regras fornecidas pelo Lingüista. Uma pilha explícita pode ser utilizada para que seja possível fazer uma comutação entre este transdutor sintático e o transdutor léxico.

O uso desses transdutores aliados aos seus mecanismos adaptativos detalhados nos capítulos anteriores garantem o processamento das regras fornecidas pelo Lingüista. O transdutor se expande, conferindo a si mesmo uma camada capaz de processar a análise sintática de uma sentença fornecida pelo usuário do sistema, conforme descrito no início deste capítulo.

As técnicas adaptativas apresentadas, são capazes de identificar e decompor todos os não-determinismos e ambigüidades presentes em uma sentença apresentada pelo usuário, segundo a gramática fornecida pelo Lingüista, criando trechos de transdutores finitos determiníticos destinados à análise da sentença em questão. Aufere-se assim um desempenho linear por partes.

Deparou-se com o fato que cada mecanismo adaptativo subjacente ao tratamento destes não-determinismos e eventualmente ambigüidades apresenta isoladamente desempenho proporcional ao comprimento da fita de entrada, ocorrências de particulares símbolos na fita de entrada, o alfabeto e o número de regras da gramática da Linguagem Natural disponíveis no sistema.

Vislumbrou-se uma sintaxe da metalinguagem para a representação da Linguagem Natural. Considerações a respeito da mesma são apresentadas no próximo capítulo, bem como a conlusão deste trabalho.