

Tecnologia Adaptativa Aplicada ao Processamento da Linguagem Natural

Ana Contier, Djalma Padovani, João José Neto

Resumo— Este trabalho faz uma breve revisão dos conceitos de Tecnologia Adaptativa, apresentando seu mecanismo de funcionamento e seus principais campos de aplicação, destacando o forte potencial de sua utilização no processamento de linguagens naturais. Em seguida são apresentados os conceitos de processamento de linguagem natural, ressaltando seu intrincado comportamento estrutural. Por fim, é apresentado o Linguístico, uma proposta de reconhecedor gramatical que utiliza autômatos adaptativos como tecnologia subjacente.

Palavras Chave— Autômatos Adaptativos, Processamento de Linguagem Natural, Reconhedores Gramaticais, Gramáticas Livre de Contexto

I. AUTÔMATOS ADAPTATIVOS

O autômato adaptativo é uma máquina de estados à qual são impostas sucessivas alterações resultantes da aplicação de ações adaptativas associadas às regras de transições executadas pelo autômato [1]. Dessa maneira, estados e transições podem ser eliminados ou incorporados ao autômato em decorrência de cada um dos passos executados durante a análise da entrada. De maneira geral, pode-se dizer que o autômato adaptativo é formado por um dispositivo convencional, não-adaptativo, e um conjunto de mecanismos adaptativos responsáveis pela auto-modificação do sistema.

O dispositivo convencional pode ser uma gramática, um autômato, ou qualquer outro dispositivo que respeite um conjunto finito de regras estáticas. Este dispositivo possui uma coleção de regras, usualmente na forma de cláusulas if-then, que testam a situação corrente em relação a uma configuração específica e levam o dispositivo à sua próxima situação. Se nenhuma regra é aplicável, uma condição de erro é reportada e a operação do dispositivo, descontinuada. Se houver uma única regra aplicável à situação corrente, a próxima situação do dispositivo é determinada pela regra em questão. Se houver mais de uma regra aderente à situação corrente do dispositivo, as diversas possíveis situações seguintes são tratadas em paralelo e o dispositivo exibirá uma operação não determinística.

Os mecanismos adaptativos são formados por três tipos de ações adaptativas elementares: consulta (inspeção do conjunto de regras que define o dispositivo), exclusão (remoção de alguma regra) e inclusão (adição de uma nova regra). As ações adaptativas de consulta permitem inspecionar o conjunto de regras que definem o dispositivo em busca de regras que sigam um padrão fornecido. As ações elementares de exclusão permitem remover qualquer regra do conjunto de regras. As ações elementares de inclusão permitem especificar a adição de uma nova regra, de acordo com um padrão fornecido.

Autômatos adaptativos apresentam forte potencial de aplicação ao processamento de linguagens naturais, devido à facilidade com que permitem representar fenômenos linguísticos complexos tais como dependências de contexto. Adicionalmente, podem ser implementados como um formalismo de reconhecimento, o que permite seu uso no pré-processamento de textos para diversos usos, tais como: análise sintática, verificação de sintaxe, processamento para traduções automáticas, interpretação de texto, corretores gramaticais e base para construção de sistemas de busca semântica e de aprendizado de línguas auxiliados por computador.

Diversos trabalhos confirmam a viabilidade prática da utilização de autômatos adaptativos para processamento da linguagem natural. É o caso, por exemplo, de [2], que mostra a utilização de autômatos adaptativos na fase de análise sintática; [3] que apresenta um método de construção de um analisador morfológico e [4], que apresenta uma proposta de autômato adaptativo para reconhecimento de anáforas pronominais segundo algoritmo de Mitkov.

II. PROCESSAMENTO DA LINGUAGEM NATURAL: REVISÃO DA LITERATURA

O processamento da linguagem natural requer o desenvolvimento de programas que sejam capazes de determinar e interpretar a estrutura das sentenças em muitos níveis de detalhe. As linguagens naturais exibem um intrincado comportamento estrutural visto que são profusos os casos particulares a serem considerados. Uma vez que as linguagens naturais nunca são formalmente projetadas, suas regras sintáticas não são nem simples nem óbvias e tornam, portanto, complexo o seu processamento computacional. Muitos métodos são empregados em sistemas de processamento de linguagem natural, adotando diferentes paradigmas, tais como métodos exatos, aproximados, pré-definidos ou interativos, inteligentes ou algorítmicos [5]. Independentemente do método utilizado, o processamento da linguagem natural envolve as operações de análise léxico-morfológica, análise sintática, análise semântica e análise pragmática [6].

A análise léxico-morfológica procura atribuir uma classificação morfológica a cada palavra da sentença, a partir das informações armazenadas no léxico [7]. O léxico ou dicionário é a estrutura de dados contendo os itens lexicais e as informações correspondentes a estes itens. Entre as informações associadas aos itens lexicais, encontram-se a categoria gramatical do item, tais como substantivo, verbo e adjetivo, e os valores morfo-sintático-semânticos, tais como gênero, número, grau, pessoa, tempo, modo, regência verbal ou nominal. Um item lexical pode ter uma ou mais representações semânticas associadas a uma entrada. É o caso da palavra “casa”, que pode aparecer das seguintes formas:

Casa: substantivo, feminino, singular, normal. Significado: moradia, habitação, sede

Casa: verbo singular, 3ª pessoa, presente indicativo, 1ª conjugação. Significado: contrair matrimônio

Dada uma determinada sentença, o analisador léxico-morfológico identifica os itens lexicais que a compõem e obtém, para cada um deles, as diferentes descrições correspondentes às entradas no léxico. A ambiguidade léxico-morfológica ocorre quando uma mesma palavra apresenta diversas categorias gramaticais. Neste caso existem duas formas de análise: a tradicional e a etiquetagem. Pela abordagem tradicional, todas as classificações devem ser apresentadas pelo analisador, deixando a resolução de ambiguidade para outras etapas do processamento. Já pela etiquetagem (POS *Tagging*), o analisador procura resolver as ambiguidades sem necessariamente passar por próximas etapas de processamento. Nesta abordagem, o analisador recebe uma cadeia de itens lexicais e um conjunto específico de etiquetas como entrada e produz um conjunto de itens lexicais com a melhor etiqueta associada a cada item. Os algoritmos para etiquetagem fundamentam-se em dois modelos mais conhecidos: os baseados em regras e os estocásticos. Os algoritmos baseados em regras usam uma base de regras para identificar a categoria de um item lexical, acrescentando novas regras à base à medida que novas situações de uso do item vão sendo encontradas. Os algoritmos baseados em métodos estocásticos costumam resolver as ambiguidades através de um corpus de treino marcado corretamente, calculando a probabilidade que uma palavra terá de receber uma etiqueta em um determinado contexto.

O passo seguinte é a análise sintática. Nesta etapa, o analisador verifica se uma sequência de palavras constitui uma frase válida da língua, reconhecendo-a ou não. O analisador sintático faz uso de um léxico e de uma gramática, que define as regras de combinação dos itens na formação das frases. A gramática adotada pode ser escrita por meio de diversos formalismos. Segundo [7] destacam-se as redes de transição, as gramáticas de constituintes imediatos (PSG ou phrase structure grammar), as gramáticas de constituintes imediatos generalizadas (GPSG) e as gramáticas de unificação funcional (PATR II e HPSG). As gramáticas de constituintes imediatos (PSG), livres de contexto, apresentam a estrutura sintática das frases em termos de seus constituintes. Por exemplo, uma frase (F) é formada pelos sintagmas nominal (SN) e verbal (SV). O sintagma nominal é um agrupamento de palavras que tem como núcleo um substantivo (Subst) e o sintagma verbal é um agrupamento de palavras que tem como núcleo um verbo. Substantivo e verbo representam classes gramaticais. O determinante (Det) compõe, junto com o substantivo, o sintagma nominal. O sintagma verbal é formado pelo verbo, seguido ou não de um sintagma nominal. O exemplo apresentado ilustra uma gramática capaz de reconhecer a frase: O menino usa o chapéu.

F → SN, SV.

SN → Det, Subst.

SV → Verbo, SN.

Det → o

Subst → menino, chapéu

Verbo → usa

No entanto, este formalismo não consegue identificar questões de concordância de gênero e número. Por exemplo, se fossem incluídos no léxico o plural e o feminino da palavra menino, frases como: “O meninos usa o chapéu.” e “O menina usa o chapéu.” seriam aceitas. Para resolver este tipo de problema existem outros formalismos, tais como o PATR II:

F → SN, SV

<SN numero> = <SV numero>

<SN pessoa> = <SV pessoa>

SN → Det, Subst

<Det numero> = <Subst numero>

<Det genero> = <Subst genero>

SV → Verbo, SN

o

<categoria> = determinante

<genero> = masc

<numero> = sing

menino

<categoria> = substantivo

<genero> = masc

<numero> = sing

chapéu

<categoria> = substantivo

<genero> = masc

<numero> = sing

usa

<categoria> = verbo

<tempo> = pres

<numero> = sing

<pessoa> = 3

<argumento 1> = SN

<argumento 2> = SN

Neste formalismo, a derivação leva em consideração outras propriedades do léxico, além da categoria gramatical, evitando os erros de reconhecimento apresentados anteriormente. Segundo [7], esse formalismo gramatical oferece poder gerativo e capacidade computacional, e tem sido usado com sucesso em ciência da computação, na especificação de linguagens de programação.

Certas aplicações necessitam lidar com a interpretação das frases bem formadas, não bastando o conhecimento da estrutura, mas sendo necessário o conhecimento do significado dessas construções. Por exemplo, quando é necessário que respostas sejam dadas a sentenças ou orações expressas em língua natural, as quais, por exemplo, provoquem um movimento no braço de um robô. Ou quando é necessário extrair conhecimentos sobre um determinado tema a partir de uma base de dados textuais. Nos casos nos quais há a necessidade de interpretar o significado de um texto, a análise léxico-morfológica e a análise sintática não são suficientes, sendo necessário realizar um novo tipo de operação, denominada análise semântica [7].

Na análise semântica procura-se mapear a estrutura sintática para o domínio da aplicação, fazendo com que a estrutura ganhe um significado [8]. O mapeamento é feito identificando as propriedades semânticas do léxico e o relacionamento semântico entre os itens que o compõe. Para representar as propriedades semânticas do léxico, pode ser usado o

formalismo PATR II, já apresentado anteriormente. Para a representação das relações entre itens do léxico pode ser usado o formalismo baseado em predicados: cada proposição é representada como uma relação predicativa constituída de um predicado, seus argumentos e eventuais modificadores. Um exemplo do uso de predicados é apresentado para ilustrar o processo de interpretação da sentença “O menino viu o homem de binóculo”. Trata-se de uma sentença ambígua da língua portuguesa, uma vez que pode ser interpretada como se (a) O menino estivesse com o binóculo, ou (b) O homem estivesse com o binóculo. Uma gramática para a análise do exemplo acima é dada pelas seguintes regras de produção:

F → SN SV
 SN → Det Subst
 SN → SN SP
 SV → V SN
 SV → V SN SP
 SP → Prep Subst

Uma possível representação semântica para as interpretações da sentença seria:

1. Sentença de interpretação (a):
 agente(ação(ver), menino)
 objeto(ação(ver), homem)
 instrumento(ação(ver), binóculo)
2. Sentença de interpretação (b):
 agente(ação(ver), menino)
 objeto(ação(ver), homem)
 qualificador(objeto(homem), binóculo)

Existem casos em que é necessário obter o conteúdo não literal de uma sentença, ligando as frases entre si, de modo a construir um todo coerente, e interpretar a mensagem transmitida de acordo com a situação e com as condições do enunciado [7]. Por exemplo, para uma compreensão literal da sentença: “O professor disse que duas semanas são o tempo necessário”, é possível recorrer aos mecanismos de representação expostos até aqui, porém para uma compreensão aprofundada, seria necessário saber a que problema se refere o professor, já que o problema deve ter sido a própria razão da formulação dessa sentença. Nestes casos, é necessária uma nova operação denominada análise pragmática.

A análise pragmática procura reinterpretar a estrutura que representa o que foi dito para determinar o que realmente se quis dizer [2]. Dois pontos focais da pragmática são: as relações entre frases e o contexto. À medida que vão sendo enunciadas, as sentenças criam um universo de referência, que se une ao já existente. A própria vizinhança das sentenças ou dos itens lexicais também constitui um elemento importante na sua interpretação. Assim, alguns novos fenômenos passam a ser estudados, como fenômenos pragmático-textuais. Inserem-se nessa categoria as relações anafóricas, co-referência, determinação, foco ou tema, dêiticos e elipse [7]. Por exemplo, nem sempre o caráter interrogativo de uma sentença expressa exatamente o caráter de solicitação de uma resposta. A sentença “Você sabe que horas são?” pode ser interpretada como uma solicitação para que as horas sejam informadas ou como uma repreensão por um atraso ocorrido. No primeiro caso, a pergunta informa ao ouvinte que o falante deseja obter uma informação e, portanto, expressa exatamente

o caráter interrogativo. Entretanto, no segundo caso, o falante utiliza o artifício interrogativo como forma de impor sua autoridade. Diferenças de interpretação desse tipo claramente implicam interpretações distintas e, portanto, problemáticas, se não for considerado o contexto de ocorrência do discurso [9]. As questões relacionadas à análise pragmática são objetos de estudos de modo a prover mecanismos de representação e de inferência adequados, e raramente aparecem em processadores de linguagem natural [7].

Em [10] são apresentados diversos artigos que mostram pesquisas atuais em processamento de linguagem natural para a Língua Portuguesa. Entre outros, apresenta-se a experiência do Núcleo Interinstitucional de Linguística Aplicada (NILC) no desenvolvimento de ferramentas para processamento de linguagem natural; o projeto VISL – Visual Interactive Syntax Learning, sediado na Universidade do Sul da Dinamarca, que engloba o desenvolvimento de analisadores morfossintáticos para diversas línguas, entre as quais o português; e o trabalho de resolução de anáforas desenvolvido pela Universidade de Santa Catarina. A tecnologia adaptativa também tem contribuído com trabalhos em processamento da linguagem natural. Em [11], são apresentadas algumas das pesquisas desenvolvidas pelo Laboratório de Linguagens e Tecnologia Adaptativa da Escola Politécnica da Universidade de São Paulo: um etiquetador morfológico, um estudo sobre processos de análise sintática, modelos para tratamento de não-determinismos e ambigüidades, e um tradutor texto-voz baseado em autômatos adaptativos.

III. RECONHECEDOR ADAPTATIVO: SUPORTE TEÓRICO LINGÜÍSTICO

Para o reconhecedor aqui proposto foi escolhida a Moderna Gramática Brasileira de Celso Luft [12] como suporte teórico linguístico, porque categoriza de forma clara e precisa os diversos tipos de sentenças de língua portuguesa, se diferenciando das demais gramáticas que priorizam a descrição da língua em detrimento da análise estrutural da mesma.

Luft diz que a oração é moldada por padrões denominados frasais ou oracionais. Estes padrões são compostos por elementos denominados sintagmas. Sintagma é qualquer constituinte imediato da oração, podendo exercer papel de sujeito, complemento (objeto direto e indireto), predicativo e adjunto adverbial. É composto por uma ou mais palavras, sendo que uma é classificada como núcleo e as demais como dependentes. As palavras dependentes podem estar localizadas à esquerda ou à direita do núcleo. Luft utiliza os seguintes nomes e abreviaturas:

1. Sintagma substantivo (SS): núcleo é um substantivo;
2. Sintagma verbal (SV): núcleo é um verbo;
3. Sintagma adjetivo (Sadj): núcleo é um adjetivo;
4. Sintagma adverbial (Sadv): núcleo é um advérbio;
5. Sintagma preposicional (SP): é formado por uma preposição (Prep) mais um SS.
6. Vlig: verbo de ligação
7. Vi: verbo intransitivo
8. Vtd: verbo transitivo direto
9. Vti: verbo transitivo indireto
10. Vtdi: verbo transitivo direto e indireto
11. Vt-pred: verbo transitivo predicativo

A Tabela 1 apresenta os elementos formadores dos sintagmas, e a sequência em que aparecem, de acordo com Luft.

TABELA 1.
Elementos formadores de sintagmas [12]

	Sintagmas
Substantivo	Quantitativos+Pronomes Adjetivos+ Sintagma Adjetivo1+Substantivo+ Sintagma Adjetivo2+ Sintagma Preposicional+ Oração Adjetiva
Verbal	Pré-verbais+ Verbo Auxiliar+ Verbo Principal
Adjetivo	Advérbio de Intensidade+ Adjetivo+ Sintagma Preposicional
Adverbial	Advérbio de Intensidade+ Adverbio+ Sintagma Preposicional
Preposicional	Preposição+ Sintagma Substantivo

Um padrão oracional é determinado pelos tipos de sintagmas e pela sequência em que aparecem. Por exemplo, o padrão oracional SS Vlig SS, indica que a frase é composta por um sintagma substantivo, seguido de um verbo de ligação e de outro sintagma substantivo. A Tabela 2 apresenta a relação de todos os padrões oracionais propostos por Luft.

Os padrões são classificados em 5 tipos:

1. Padrões pessoais nominais: Neste caso, existe sujeito e o núcleo do predicado é um nome (substantivo, adjetivo, advérbio) ou um pronome (substantivo, adjetivo, advérbio). O verbo, nesses casos, é chamado de verbo de ligação (Vlig).

TABELA 2
Padrões oracionais de Luft [12]

Padrões Pessoais Nominais				
SS	Vlig	SS		
SS	Vlig	Sadj		
SS	Vlig	Sadv		
SS	Vlig	SP		
Padrões Pessoais Verbais				
SS	Vtd	SS		
SS	Vti	SP		
SS	Vti	Sadv		
SS	Vti	SP	SP	
SS	Vtdi	SS	SP	
SS	Vtdi	SS	Sadv	
SS	Vtdi	SS	SP	SP
SS	Vi			

2. Padrões pessoais verbais: São aqueles nos quais existe o sujeito e o núcleo do predicado é um verbo. O verbo pode ser transitivo direto (Vtd), transitivo indireto (Vti), transitivo direto e indireto (Vtdi), e intransitivo (Vi). Se o verbo for

transitivo direto (Vtd), o complemento será um objeto direto; se o verbo for transitivo indireto (Vti), o complemento será um objeto indireto; se o verbo for transitivo direto e indireto (Vtdi), o complemento será um objeto direto e um indireto; se o verbo for intransitivo (Vi), não há complemento.

TABELA 2 - CONTINUAÇÃO

Padrões Pessoais Verbo-Nominais			
SS	Vtpred	SS	SS
SS	Vtpred	SS	Sadj
SS	Vtpred	SS	SP
SS	Vtpred	SS	Sadv
SS	Vtpred	SS	
SS	Vtpred	Sadj	
SS	Vtpred	SP	
Padrões Impessoais Nominais			
	Vlig	SS	
	Vlig	Sadj	
	Vlig	Sadv	
	Vlig	SP	
Padrões Impessoais Verbais			
	Vtd	SS	
	Vti	SP	
	Vi		

3. Padrões Pessoais Verbo-Nominais: Neste caso, existe o sujeito e o núcleo do predicado é um verbo transitivo predicativo (Vt-pred), cujo complemento é um objeto direto e um predicativo do objeto.

4. Padrões Impessoais Nominais: Ocorrem quando não existe sujeito e o núcleo do predicado é um nome (substantivo, adjetivo, advérbio) ou um pronome (substantivo, adjetivo, advérbio).

5. Padrões Impessoais Verbais: Neste caso, não existe sujeito e o núcleo do predicado é um verbo.

Luft apresenta uma gramática usada para análise sintática da Língua Portuguesa no modelo moderno, em que as frases são segmentadas o mais binariamente possível: Sujeito+Predicado; Verbo+Complemento; Substantivo+Adjetivo, etc. Neste modelo, a descrição explicita somente as classes analisadas; as funções ficam implícitas. Querendo explicar estas, Luft sugere que sejam escritas à direita das classes: SS:Sj (Sujeito), V:Núc (Núcleo do Predicado), PrA:NA (Adjunto Adnominal), etc. A gramática proposta por Luft é a seguinte:

F → [Conec] [SS] SV [Conec]
 Conec → F
 SS → [Sadj] SS [Sadj | SP]
 SS → [Quant | PrA] (Sc | Sp | PrPes)
 SV → [Neg] [Aux | PreV] (Vlig | Vtd | Vti | Vtdi | Vi)
 [SS | Sadj | Sadv | SP] [SSI | Sadj | Sadv | SP] [SP]
 SP → Prep (SS | Sadj)
 Sadj → Sadj [SP]
 Sadj → [Adv] Adj

Sadv → Sadv [SP]

Sadv → [Adv] Adv

PrA → Ind | ArtDef | ArtInd | Dem | Pos

Sendo:

F – Frase

SS – Sintagma substantivo

SV – Sintagma verbal

SP – Sintagma preposicional

SN – Sintagma nominal

Sadv – Sintagma adverbial

Sadj – Sintagma adjetivo

Adv – advérbio

Adj – adjetivo

ArtDef – artigo definido

ArtInd – artigo indefinido

Aux – Partícula auxiliar (apassivadora ou pré-verbal)

Conec – Conector (conjunção ou pronome relativo)

Dem – pronome demonstrativo indefinido

Ind – pronome indefinido

Neg – partícula (negação)

PrA – pronome adjetivo

PrPes – pronome pessoal

Prep – preposição

Quant – numeral

Sc – substantivo comum

Sp – substantivo próprio

V – verbo

Vlig – verbo de ligação

Vi – verbo intransitivo

Vtd – verbo transitivo direto

Vti – verbo transitivo indireto

Vtdi – verbo transitivo direto e indireto

IV. PROPOSTA DE UM RECONHECEDOR GRAMATICAL

O Linguístico é uma proposta de reconhecedor gramatical composto de 5 módulos sequenciais que realizam cada qual um processamento especializado, enviando o resultado obtido para o módulo seguinte, tal como ocorre em uma linha de produção, até que o texto esteja completamente analisado.

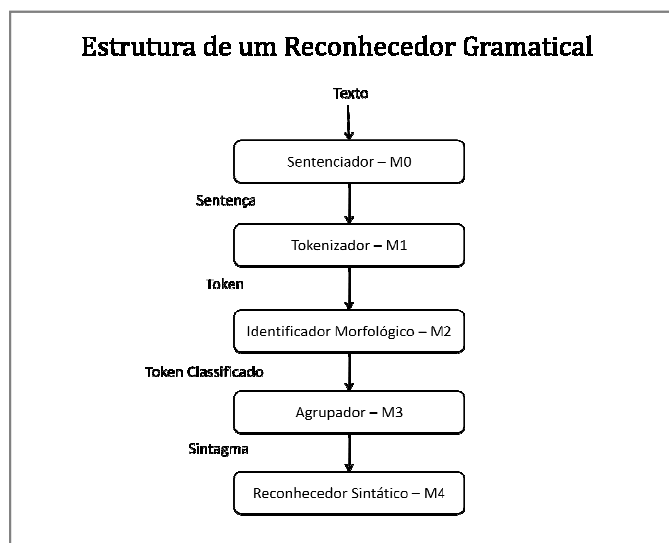


Figura 2. Estrutura para a proposta de um Reconhecedor Gramatical

A Fig.2 ilustra a estrutura do Linguístico. O primeiro módulo, denominado Sentenciador, recebe um texto e realiza um pré-processamento, identificando os caracteres que possam indicar final de sentença, eliminando hífen, apóstrofes e aspas, e substituindo vírgulas e ponto-e-vírgulas por caracteres especiais. Ao final, o Sentenciador divide o texto em supostas sentenças que são analisadas individualmente nas etapas seguintes.

O segundo módulo, denominado Tokenizador, recebe as sentenças identificadas na etapa anterior e as divide em *tokens*, considerando, neste processo, abreviaturas, valores monetários, numerais arábicos e romanos, nomes próprios, caracteres especiais e de pontuação final. Os *tokens* são armazenados em estruturas de dados (*arrays*) e enviados um a um para análise do módulo seguinte.

O terceiro módulo, denominado Identificador Morfológico, recebe os *tokens* da etapa anterior e os identifica morfológicamente, utilizando os textos pré-annotados do corpus Bosque[13] como biblioteca de apoio. O Bosque é um conjunto de frases anotadas morfossintaticamente (conhecido por *treebank*), composto por 9368 frases retiradas dos primeiros 1000 extratos dos corpora CETEMPublico (Corpus de Extractos de Textos Electrónicos MCT/Público) e CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo). A Fig. 3 apresenta um fragmento do Bosque. Como o Bosque pode apresentar mais de uma classificação morfológica para o mesmo *token*, o Identificador Morfológico usa aquela que é mais frequente, identificando artigos, numerais, substantivos, nomes próprios, verbos, advérbios, adjetivos, preposições e conjunções. Em seguida, o Identificador usa um conjunto de regras heurísticas resultantes da análise prévia de textos de treinamento para resolver situações nas quais não foi encontrada uma classificação ou a classificação encontrada não é a mais adequada no contexto de análise, que é identificado pela classificação do último *token* analisado. Por exemplo, se não for encontrada classificação morfológica para um *token*, ele terminará em ‘ado’, ou ‘ido’, e a classificação do último *token* for verbo, então o Identificador Morfológico o classifica como verbo.

```

#9363 CF997-3 Segundo declarações do próprio diretor, ele vive até hoje de forma angustiada:
[STA+fcl [ADVL+pp
    (H+prp Segundo)
    (P<+np
    (H+n declarações)
    (N<+pp
    (H+prp de)
    (P<+np
    (N+art o)
    (N+pron-det próprio)
    (H+n diretor)))]
    (.)
    (SUBJ+pron-pers ele)
    (P+v-fin vive)
    (ADVL+advp
    (A+prp até)
    (H+adv hoje))
    (ADVL+pp
    (H+prp de)
    (P<+np
    (H+n forma)
    (N<+v-ppc angustiada))]
    (.)
  
```

Figura 3. Exemplo de frase etiquetada do corpus Bosque [11].

O quarto módulo, denominado Agrupador é composto de um autômato, responsável pela montagem dos sintagmas a partir

de símbolos terminais da gramática e um bigrama, responsável pela montagem dos sintagmas a partir de não-terminais. Inicialmente, o Agrupador recebe do Identificador as classificações morfológicas dos *tokens* e as agrupa em sintagmas de acordo com a gramática proposta por Luft. Neste processo são identificados sintagmas nominais, verbais, preposicionais, adjetivos e adverbiais. Para isso, o Agrupador utiliza um autômato adaptativo cuja configuração completa é definida da seguinte forma:

Estados = {1, 2, 3, 4, SS, SP, V, Sadj, Sadv, A},

Onde:

1,2,3 e 4 = Estados Intermediários

SS, SP, V Sadj, Sadv = Estados nos quais houve formação de sintagmas, sendo:

SS= Sintagma substantivo

SP = Sintagma preposicional

V = Verbo ou locução verbal

Sadj = Sintagma adjetivo

Sadv = Sintagma adverbial

A = Estado após o processamento de um ponto final

Tokens = {art, num, n, v, prp, pron, conj, adj, adv, rel, pFinal, sClass}, onde:

art = artigo, num = numeral

n = substantivo, v = verbo

prp = preposição, pron = pronome

conj = conjunção, adj = adjetivo

adv = advérbio, rel = pronome relativo

pFinal = ponto final, sClass = sem classificação

Estados de Aceitação = {SS, SP, V, Sadj, Sadv, A}

Estado Inicial = {1}

Função de Transição = {(Estado, *Token*)→Estado}, sendo:

{(1, art)→2, (2, art)→2, (3, art)→3

(1, num)→Sadv, (2, num)→2, (3, num)→3

(1, n)→SS, (2, n)→SS, (3, n)→SP

(1, v)→SV, (2, v)→SV, (3, v)→SP

(1, prp)→3, (2, prp)→2, (3, prp)→3

(1, prop)→SS, (2, prop)→SS, (3, prop)→SP

(1, pron)→SS, (2, pron)→SS, (3, pron)→SP

(1, conj)→conj, (2, conj)→∅, (3, conj)→∅

(1, adj)→Sadj, (2, adj)→Sadj, (3, adj)→3

(1, adv)→Sadv, (2, adv)→2, (3, adv)→3

(1, rel)→conj, (2, rel)→∅, (3, rel)→conj

(1, pFinal)→A, (2, pFinal)→∅, (3, pFinal)→∅}

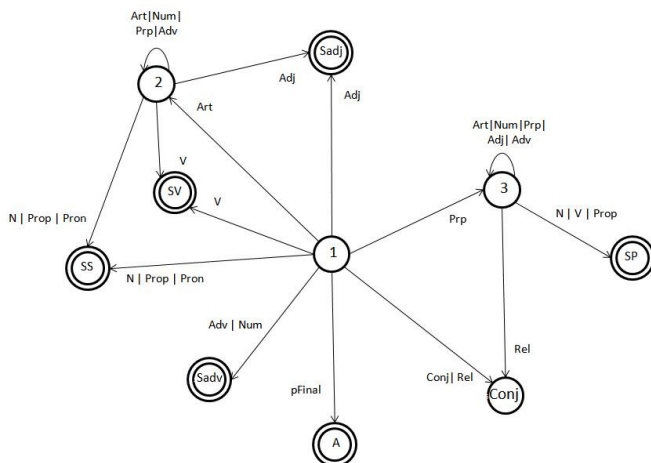


Figura 4. Configuração Completa do Autômato Construtor de Sintagmas.

Por exemplo, segundo a gramática de Luft, os sintagmas substantivos são obtidos através da seguinte regra:

SS → [Quant | PrA] (Sc | Sp | PrPes)

Pela regra acima, o conjunto de *tokens* “A” e “casa” formam um sintagma substantivo, da seguinte forma:

PrA = “A” (artigo definido)

Sc = “casa” (substantivo comum)

Da direita para esquerda, são realizadas as seguintes derivações:

PrA Sc → SS; A Sc → SS; A casa → SS

Já o Agrupador recebe o *token* “A”, identificado pelo Tokenizador como artigo definido, e se movimenta do estado 1 para o estado 2. Ao receber o *token* “casa”, identificado como substantivo comum, ele se movimenta do estado 2 para o estado SS, que é um estado de aceitação. Neste momento o Agrupador armazena a cadeia “A casa” e o símbolo “SS” em uma pilha e reinicializa o autômato preparando-o para um novo reconhecimento.

Em um passo seguinte, o Agrupador usa o bigrama para comparar um novo sintagma com o último sintagma formado, visando identificar elementos mais altos na hierarquia da gramática de Luft. Para isso ele usa a matriz apresentada na Tabela 3, construída a partir da gramática de Luft. A primeira coluna da matriz indica o último sintagma formado (US) e a primeira linha, o sintagma atual (SA). A célula resultante apresenta o novo nó na hierarquia da gramática.

TABELA 3
Matriz de agrupamento de sintagmas

SA \ US	SS	SP	V	Sadv	Sadj	Conj
SS	SS	SS	-	-	SS	-
SP	SP	-	-	-	-	-
V	-	-	V	-	-	-
Sadv	-	-	-	Sadv	Sadj	-
Sadj	SS	Sadj	-	-	Sadj	-
Conj	-	-	-	-	-	Conj

Esta técnica foi usada para tratar as regras gramaticais nas quais um sintagma é gerado a partir da combinação de outros, como é o caso da regra de formação de sintagmas substantivos: SS → [Sadj] SS [Sadj | SP]. Por esta regra, os sintagmas substantivos são formados por outros sintagmas substantivos precedidos de um sintagma adjetivo ou seguidos de um sintagma adjetivo ou um sintagma preposicional. No exemplo anterior, supondo que os próximos 2 *tokens* fossem “de” e “madeira”, após a passagem pelo autômato, o Agrupador formaria um sintagma SP. Considerando que na pilha ele tinha armazenado um SS, após a passagem pelo bigrama, e de acordo com a Tabela 3, o sintagma resultante seria um SS e o conteúdo que o compõe seria a combinação dos textos de cada sintagma que o originou. Caso não haja agrupamentos possíveis, o Agrupador envia o último sintagma formado para análise do Reconhecedor Sintático e movimenta

o sintagma atual para a posição de último sintagma no bigrama, repetindo o processo com o próximo sintagma.

O quinto e último módulo, denominado Reconhecedor Sintático, recebe os sintagmas do módulo anterior e verifica se estão sintaticamente corretos de acordo com padrões gramaticais de Luft. O Reconhecedor Sintático utiliza um autômato adaptativo que faz chamadas recursivas sempre que recebe conjunções ou pronomes relativos, armazenando, em uma estrutura de pilha, o estado e a cadeia de sintagmas reconhecidos até o momento da chamada. Caso o Reconhecedor Sintático não consiga se movimentar a partir do sintagma recebido, ele gera um erro e retorna o ponteiro para o último sintagma reconhecido, finalizando a instância do autômato recursivo e retornando o processamento para aquela que a inicializou. Esta, por sua vez, retoma posição em que se encontrava antes da chamada e continua o processamento até o final da sentença ou até encontrar uma nova conjunção, situação na qual o processo se repete.

A configuração completa do autômato é definida da seguinte forma:

Estados = { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27 }

Tokens = { SS, SP, Vli, Vi, Vtd, Vti, Vtdi, Sadj, Sadv, Conj, A }

Estados de Aceitação = { 4, 5, 6, 9, 12, 13, 14, 15, 17, 18, 19, 21, 22, 24, 25, 26, 27 }

Estado Inicial = { 1 }

Função de Transição = { (Estado, Token) → Estado }, sendo:
 { (1, SS) → 2, (2, Vti) → 3, (3, SP) → 4, (4, SP) → 4,
 (3, Sadv) → 5, (2, Vi) → 6, (2, Vtdi) → 7
 (7, SS) → 8, (8, SP) → 9, (9, SP) → 9, (8, Sadv) → 10,
 (2, Vlig) → 11, (11, SP) → 12, (11, Sadv) → 13,
 (11, Sadj) → 14, (11, SS) → 15, (2, Vtd) → 16, (16, SS) → 17,
 (2, Vtpred) → 18, (18, SP) → 19, (18, Sadj) → 20
 (18, SS) → 21, (21, SS) → 22, (21, Sadj) → 23, (21, Sadv) → 24,
 (21, SP) → 25 }

Pilha = { [Texto, Sintagma, Estado] }

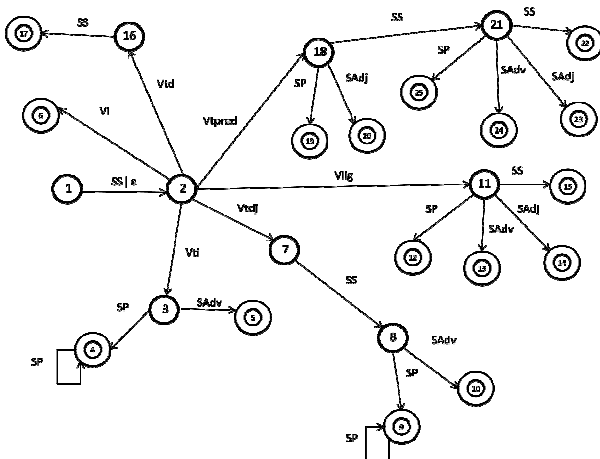


Figura 5.1. Configuração Completa do Reconhecedor Sintático.

No entanto, para que a análise sintática seja feita, não são necessárias todas as ramificações da configuração completa do autômato. Por exemplo, quando se transita um verbo de ligação a partir do estado 2, o autômato vai para o estado 11 e todas as demais ramificações que partem deste estado para os estados 3, 7, 16 e 18, não são usadas. Com a tecnologia adaptativa, é possível criar dinamicamente os estados e transições do autômato em função dos tipos de verbos, evitando manter ramificações que não são usadas.

A Fig. 5.2 apresenta a configuração inicial do autômato adaptativo equivalente ao autômato de pilha apresentado anteriormente. No estado 1, o autômato recebe os *tokens* e transita para o estado 2 quando processa um sintagma substantivo (SS) ou quando transita em vazio. No estado 2, o autômato transita para si mesmo quando recebe qualquer tipo de verbo: Vi, Vtd, Vlig, Vtpred, Vtdi e Vti. Todas as outras ramificações são criadas por meio de funções adaptativas chamadas em função do tipo de verbo processado.

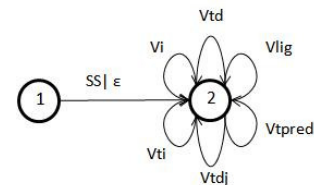


Figura 5.2. Configuração Inicial do Reconhecedor Gramatical.

Por exemplo, se o verbo é de ligação (Vlig), o autômato utiliza as funções adaptativas $\alpha(j)$ e $\beta(o)$, definidas da seguinte forma:

$\alpha(j): \{ o^* : \begin{aligned} & - [(j, Vlig)] \\ & + [(j, Vlig) : \rightarrow o, \beta(o)] \end{aligned} \}$

$\beta(o): \{ t^*u^*v^*x^* : \begin{aligned} & + [(o, SP) : \rightarrow t] \\ & + [(o, Sadv) : \rightarrow u] \\ & + [(o, Sadj) : \rightarrow v] \\ & + [(o, SS) : \rightarrow x] \end{aligned} \}$

A função adaptativa $\alpha(j)$ é chamada pelo autômato antes de processar o *token*, criando o estado 11 e a produção que leva o autômato do estado 2 ao novo estado criado. Em seguida, o autômato chama a função $\beta(o)$, criando os estados 12, 13, 14 e 15 e as produções que interligam o estado 11 aos novos estados. A Fig. 5.3 mostra a configuração do autômato após o processamento do verbo de ligação. Neste exemplo, o autômato criou apenas os estados 11, 12, 13, 14 e 15 e as respectivas transições, evitando alocar recursos que seriam necessários para criar o autômato completo, conforme apresentado na Fig. 5.1.

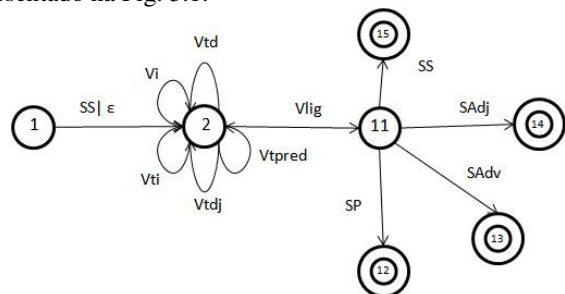


Figura 5.3. Configuração do autômato após o processamento do verbo de ligação.

Toda movimentação do autômato, assim como os sintagmas identificados em cada passagem e a classificação morfológica dos termos das sentenças, são armazenados em arquivos que podem ser acessados por um editor. Se o Linguístico não consegue reconhecer a sentença, ele registra os erros encontrados e grava uma mensagem alertando para o ocorrido.

V. EXPERIMENTO

Foi realizado um experimento no qual a implementação conceitual descrita na seção anterior foi codificada. A linguagem de programação utilizada foi Python [14] por ela dispor de recursos de processamento de expressões regulares (*Regular Expression* – RE), processamento de linguagem natural (*Natural Language Tool Kit* – NLTK) e autômatos finitos (*Determinist Finite Automon* – DFA) necessários para a realização do trabalho. Segue um exemplo de texto analisado pelo reconhecedor e os resultados obtidos:

Texto:

“A Câmara de Representantes do Congresso americano deu início ao debate de um projeto de lei histórico para limitar as emissões de gases causadores do efeito estufa pela indústria dos Estados Unidos.”

Resultados obtidos:

Classificação Morfológica:

['A', 'art'], ['Câmara', 'prop'], ['de', 'prp'],
['Representantes', 'prop'], ['do', 'prp'], ['Congresso', 'prop'],
['americano', 'adj'], ['deu', 'v'], ['início', 'n'],
['ao', 'prp'], ['debate', 'n'], ['de', 'prp'], ['um', 'art'],
['projeto', 'n'], ['de', 'prp'], ['lei', 'n'], ['histórico', 'adj'],
['para', 'prp'], ['limitar', 'v'], ['as', 'art'], ['emissões', 'n'],
['de', 'prp'], ['gases', 'n'], ['causadores', 'adj'], ['do', 'prp'],
['efeito', 'n'], ['estufa', 'adj'], ['pela', 'prp'], ['indústria', 'n'],
['dos', 'prp'], ['Estados', 'prop'], ['Unidos', 'prop']]

Total de Sintagmas= 5

[[['A Câmara', 'SS', '2'],
['de Representantes do Congresso americano', 'SP', '2'],
['deu', 'V', '3'],
['início', 'SS', '4'],
['ao debate de um projeto de lei histórico para limitar as emissões de gases causadores do efeito estufa pela indústria dos Estados Unidos', 'SP', '9']]]

VI. CONSIDERAÇÕES FINAIS

Este artigo apresentou uma revisão dos conceitos de Tecnologia Adaptativa e de Processamento da Linguagem Natural. Em seguida, foi apresentado o Linguístico, uma proposta de reconhecedor gramatical que utiliza autômatos adaptativos como tecnologia subjacente. Ao final, foi

apresentado um experimento realizado com o reconhecedor e os resultados obtidos.

IX. REFERÊNCIAS

- [1] NETO, J.J. Apresentação LTA-Laboratório de Linguagens e Técnicas Adaptativas. Disponível em: <http://www.pcs.usp.br/~lta>. Acesso 01/11/2009.
- [2] TANIWAKI, C. Formalismos adaptativos na análise sintática de linguagem natural. Dissertação de Mestrado, EPUSP, São Paulo, 2001.
- [3] MENEZES, C. E. Um método para a construção de analisadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos. Dissertação de Mestrado, Escola Politécnica da Universidade de São Paulo, 2000.
- [4] PADOVANI, D. Uma proposta de autômato adaptativo para reconhecimento de anáforas pronominais segundo algoritmo de Mitkov. Workshop de Tecnologias Adaptativas – WTA 2009, 2009.
- [5] MORAES, M. de Alguns aspectos de tratamento sintático de dependência de contexto em linguagem natural empregando tecnologia adaptativa, Tese de Doutorado, Escola Politécnica da Universidade de São Paulo, 2006.
- [6] RICH, E.; KNIGHT, K. Inteligência Artificial, 2. Ed. São Paulo: Makron Books, 1993.
- [7] VIEIRA, R.; LIMA, V. Linguística computacional: princípios e aplicações. IX Escola de Informática da SBC-Sul, 2001.
- [8] FUCHS, C., LE GOFFIC, P. Les Linguistiques Contemporaines.
- [9] NUNES, M. G. V. et al. Introdução ao Processamento das Línguas Naturais. Notas didáticas do ICMC N° 38, São Carlos, 88p, 1999.
- [13] Paris, Hachette, 1992. 158p.
- [10] SARDINHA, T. B. A Língua Portuguesa no Computador. 295p. Mercado de Letras, 2005.
- [11] ROCHA, R.L.A. Tecnologia Adaptativa Aplicada ao Processamento Computacional de Língua Natural. Workshop de Tecnologias Adaptativas – WTA 2007, 2007.
- [12] LUFT, C. Moderna Gramática Brasileira. 2ª. Edição Revista e Atualizada. 265p. Editora Globo, 2002.
- [13] Linguatca : <http://www.linguatca.pt/>
- [14] Python. Python Programming Language – Official Website. Disponível em: <<http://www.python.org>>.

Ana Teresa Contier: formada em Letras-Português pela Universidade de São Paulo (2001) e em publicidade pela PUC-SP (2002). Em 2007 obteve o título de mestre pela Poli-USP com a dissertação: “Um modelo de extração de propriedades de textos usando pensamento narrativo e paradigmático”.

Djalma Padovani nasceu em São Paulo em 1964. cursou bacharelado em Física pelo Instituto de Física da Universidade de São Paulo, formou-se em administração de empresas pela Faculdade de Economia e Administração da Universidade de São Paulo, em 1987 e obteve o mestrado em engenharia de software pelo Instituto de Pesquisas Tecnológicas de São Paulo - IPT, em 2008. Trabalhou em diversas empresas nas áreas de desenvolvimento de software e tecnologia de informação e atualmente é responsável pela arquitetura tecnológica da Serasa S/A, empresa do grupo Experian.

João José Neto graduado em Engenharia de Eletricidade (1971), mestrado em Engenharia Elétrica (1975) e doutorado em Engenharia Elétrica (1980), e livre-docência (1993) pela Escola Politécnica da Universidade de São Paulo.

Atualmente é professor associado da Escola Politécnica da Universidade de São Paulo, e coordena o LTA - Laboratório de Linguagens e Tecnologia Adaptativa do PCS - Departamento de Engenharia de Computação e Sistemas Digitais da EPUSP. Tem experiência na área de Ciência da Computação, com ênfase nos Fundamentos da Engenharia da Computação, atuando principalmente nos seguintes temas: dispositivos adaptativos, tecnologia adaptativa, autômatos adaptativos, e em suas aplicações à Engenharia de Computação, particularmente em sistemas de tomada de decisão adaptativa, análise e processamento de linguagens naturais, construção de compiladores, robótica, ensino assistido por computador, modelagem de sistemas inteligentes, processos de aprendizagem automática e inferências baseadas em tecnologia adaptativa.