

Linguístico: Usando Tecnologia Adaptativa para a Construção de Um Reconhecedor Gramatical

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, Reconhecedores Gramaticais, Gramáticas Livres de Contexto

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.

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 simples nem tampouco 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

Considerando o processamento da direita para esquerda e de baixo para cima, a frase seria analisada da seguinte forma:

O menino usa o chapéu

1. chapéu = Subst: O menino usa o subst

2. O = Det : O menino usa det subst

3. Det Subst = SN:O menino usa SN

4. usa = verbo: O menino verbo SN

5. verbo SN=SV: O menino SV

6. menino = Subst: O subst SV

7. O = Det: Det subst SV

8. Det subst= SN: SN SV

9. SN SV= F: F (Aceita)

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. Por exemplo:

O meninos usa o chapéu

1. chapéu = Subst: O menino usa o subst

2. O = Det : O menino usa det subst

3. Det Subst = SN:O menino usa SN

4. usa = verbo: O menino verbo SN

5. verbo SN=SV: O menino SV

6. meninos = Subst: O subst SV

7. O = Det: Det subst SV

8. Det subst= SN: SN SV

9. SN SV= F: F (Aceita)

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. Aplicando este formalismo ao exemplo acima, o erro de concordância seria identificado e a frase não seria aceita:

O meninos usa o chapéu

1. chapéu = Subst masc sing : O menino usa o subst
2. O = Det : O menino usa det subst
3. Det Subst = SN:O menino usa SN
4. usa = verbo pres 3a. Pessoa sing: O menino verbo SN
5. verbo SN=SVsing: O menino SVsing
6. meninos = Subst masc plural: O subst SVsing
7. O = Det: Det subst SVsing
8. Det subst= SNplur: SNplur SVsing
9. SNplur SVsing = Não aceita

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, provocam 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:

- $$\begin{aligned}
 F &\rightarrow SN\ SV \\
 SN &\rightarrow Det\ Subst \\
 SN &\rightarrow SN\ SP \\
 SV &\rightarrow V\ SN \\
 SV &\rightarrow V\ SN\ SP \\
 SP &\rightarrow Prep\ Subst
 \end{aligned}$$

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 trabalhos de pesquisas em processamento de linguagem natural para a Língua Portuguesa tais como o desenvolvido pelo 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 ambiguidades, e um tradutor texto voz baseado em autômatos adaptativos.

RECONHECEDOR ADAPTATIVO: SUPORTE TEÓRICO LINGUÍSTICO

A Moderna Gramática Brasileira de Celso Luft [12] foi escolhida como suporte teórico linguístico do reconhecedor aqui proposto. A escolha foi feita em função da forma clara e precisa com que Luft categoriza 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		

TABELA 2 - CONTINUAÇÃO

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 explícita 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] [SS | 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

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.

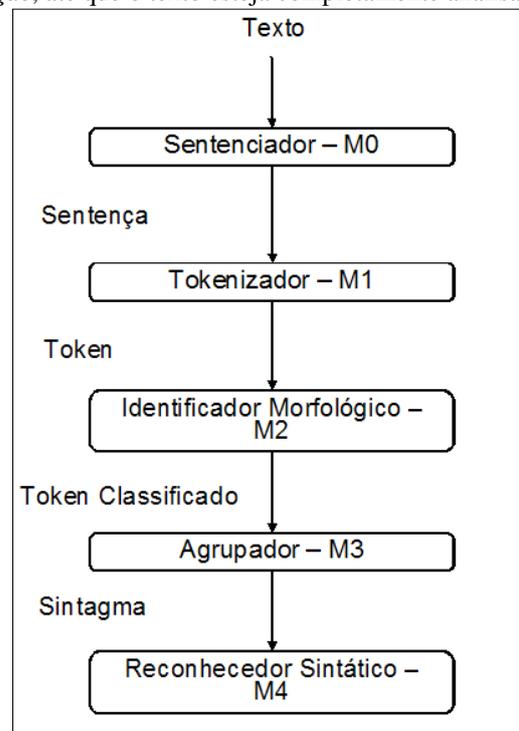


Figura 2. Estrutura do Linguístico

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, palavras abreviadas e palavras compostas, e eliminando aspas simples e duplas. Ao final, o Sentenciador divide o texto em supostas sentenças, para análise individual 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, horas e minutos, numerais arábicos e romanos, palavras compostas, 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, como biblioteca de apoio, os

textos pré-annotados do corpus Bosque[13], os verbos, substantivos e adjetivos que fazem parte da base de dados do TeP2.0 – Thesouro Eletrônico para o Português do Brasil [14] e as conjunções, preposições e pronomes disponíveis no Portal São Francisco[15], cujas informações provêm da Wikipedia [16]. O Bosque é um conjunto de frases anotadas morfossintaticamente (conhecido por *treebank*), composto por 9368 frases retiradas dos primeiros 1000 extratos do corpora CETEMPúblico (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.

```
#5156 CP1003-14 Todas as interpretações são possíveis, tudo pode
acontecer a este corpo.
(STA+cu (CJT+fcl (SUBJ+np
    (>N+pron-det Todas as)
    (H+n interpretações))
    (P+v-fin são)
    (SC+adj possíveis))
    (,))
    (CJT+fcl (SUBJ+pron-indp tudo)
    (P+vp
    (AUX+v-fin pode)
    (MV+v-inf acontecer))
    (PIV+pp
    (H+prp a)
    (P<+np
    (>N+pron-det este)
    (H+n corpo))))
    (,))
```

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

O TeP2.0 é um dicionário eletrônico de sinônimos e antônimos para o português do Brasil, que armazena conjuntos de formas léxicas sinônimas e antônimas. É composto por 19.888 conjuntos de sinônimos, 44.678 unidades lexicais, e 4.276 relações de antonímia [14].

O Portal São Francisco apresenta um curso online da Língua Portuguesa e, entre seus módulos, encontra-se um sumário das classes morfológicas, no qual são encontrados exemplos de palavras e locuções mais comuns de cada classe [15].

Inicialmente, o Identificador Morfológico procura pela classificação morfológica dos tokens no léxico do Bosque, caso não a encontre, então ele procura na base de dados do TeP2.0 e no léxico do Portal São Francisco.

O padrão de etiquetas usado pelo Linguístico é o mesmo do Bosque, que apresenta, além da classificação morfológica, o papel sintático que o token exerce na sentença pré-annotada. Por exemplo, a etiqueta >N+art indica que o token é um artigo que está à esquerda de um substantivo (>N). A notação usa como gramática subjacente a Gramática Constitutiva proposta por Fred Karlsson [17].

O léxico do Bosque foi organizado de modo a relacionar todas as classificações de um tokens ordenadas por frequência em que ocorrem no texto pré-annotado. Por exemplo, o token “acentuada” está classificado com as seguintes etiquetas: N<+v-pp e P+v-pp, o que significa que, no texto pré-annotado, ele foi classificado como verbo no particípio (+v-pp) antecedido de um substantivo (N<) e como verbo no particípio no papel de predicador (P).

No caso de ambiguidade, o Identificador Morfológico assume a classificação mais frequente como inicial e verifica se a classificação mais frequente do token seguinte é consistente com o que indica a etiqueta do token analisado. Se

for, vale a classificação mais frequente, senão o Identificador analisa a próxima classificação, repetindo o algoritmo. Caso o algoritmo não retorne uma classificação única, o Identificador passa todas as classificações encontradas para os módulos seguintes, para que ambiguidade seja resolvida pelas regras gramaticais do reconhecedor.

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 (Fig.4). 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)→∅}

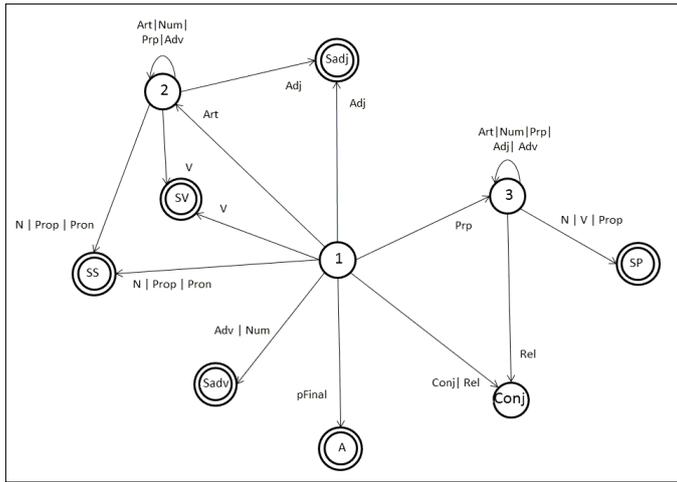


Figura 4. Configuração Completa do Autômato Construtor de Sintagmas.

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

$$SS \rightarrow [Quant | PrA] (Sc | Sp | PrPes)$$

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 transições:

$$PrA Sc \rightarrow SS; A Sc \rightarrow SS; A casa \rightarrow SS$$

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: $SS \rightarrow [Sadj] SS [Sadj | SP]$. Por esta regra, os sintagmas substantivos são formados por outros sintagmas substantivos precedidos de um sintagma adjetivo e 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] }

submáquinas especialistas. O Autômato Mestre é responsável pelo sequenciamento das chamadas às submáquinas, de acordo com um conjunto de regras cadastradas em base de dados.

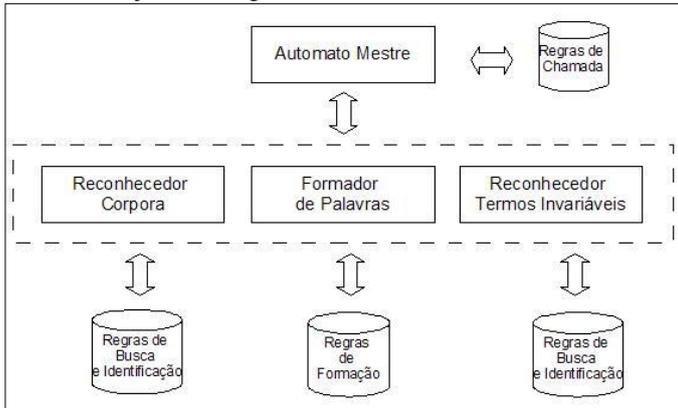
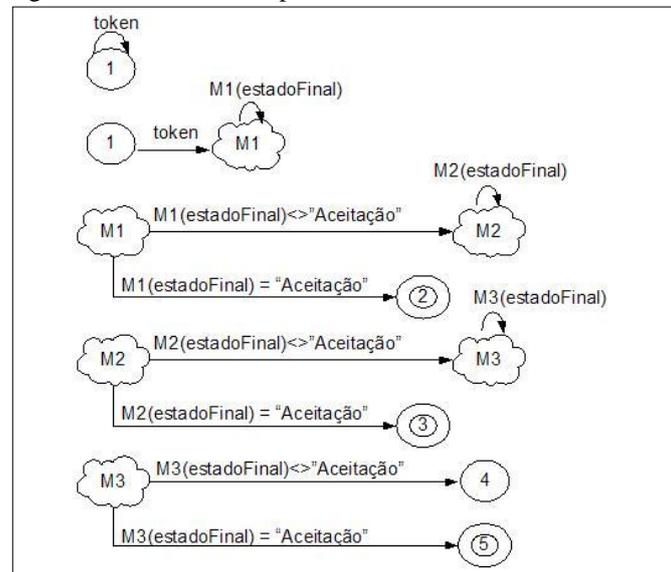


Figura 6.1. Arquitetura do Identificador Morfológico.

Inicialmente, o Autômato Mestre busca as classificações morfológicas no c (no caso, o Bosque, mas poderia ser outro, se houvesse necessidade); em seguida procura por substantivos, adjetivos e verbos, no formato finito e infinito (flexionados e não flexionados), através de uma submáquina de formação e identificação de palavras, que usa, como base, o vocabulário do TeP2.0 (aqui também existe a possibilidade de usar outra base de dados, substituindo ou complementando o TeP2.0); por fim, o Autômato Mestre procura por termos invariáveis, ou seja, termos cuja classificação morfológica é considerada estável pelos linguistas, tais como, conjunções, preposições e pronomes, no caso, extraídos no léxico do Portal São Francisco. A Fig. 6.2 apresenta a estrutura adaptativa do Autômato Mestre.

Figura 6.2. Estrutura Adaptativa do Autômato Mestre.



O Autômato Mestre inicia seu processamento recebendo o token da coleção de tokens criada pela classe Texto. Em seguida, antes de processá-lo, o autômato se modifica através de uma função adaptativa, criando uma submáquina de processamento (M1), uma transição entre o estado 1 e M1, e uma transição que aguarda o estado final da submáquina M1. O token é passado para M1 e armazenado em uma pilha. Quando M1 chega ao estado final, o autômato se modifica novamente, criando uma nova submáquina M2, o estado 2 e as

transições correspondentes. Caso o estado final de M1 seja de aceitação, o processo é finalizado no estado 2, caso contrário a máquina M2 é chamada, passando o token armazenado na pilha. O processo se repete quando M2 chega ao final do processamento, com a criação da submáquina M3, do estado 3 e das transições correspondentes. Um novo ciclo se repete e se o estado final de M3 é de aceitação, o autômato transiciona para o estado 5, de aceitação, caso contrário, ele vai para o estado 4 de não aceitação. M1, M2 e M3 representam, respectivamente, as submáquinas do Reconhecedor de Corpus, do Formador de Palavras e do Reconhecedor de Termos Invariáveis. Portanto, caso o Autômato Mestre encontre a classificação morfológica ao final de M1, ele não chama M2; caso encontre em M2, não chama M3 e, caso também não encontre em M3, ele informa aos demais módulos do Linguístico que não há classificação morfológica para o termo analisado.

As submáquinas M1, M2 e M3 também foram projetadas de acordo com a tecnologia adaptativa. A Máquina M1 usa um autômato adaptativo que se automodifica de acordo com o tipo de token que está sendo analisado: palavras simples, palavras compostas, números, valores e símbolos. A Fig. 6.3 apresenta a estrutura adaptativa de M1.

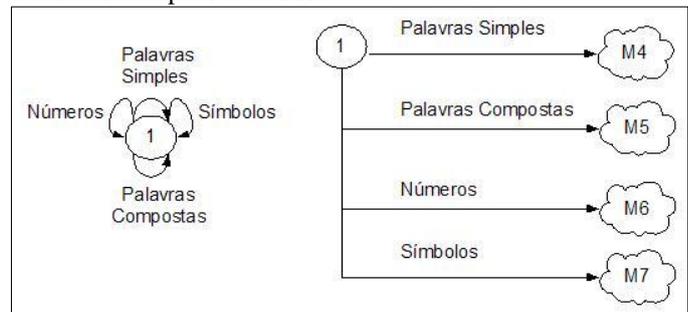


Figura 6.3. Estrutura Adaptativa do Reconhecedor de Corpus M1.

Inicialmente o autômato é composto por um único estado e por transições para ele mesmo (Fig.6.3, à esquerda). Ao identificar o tipo de token que será analisado (obtido no processo de tokenização), o autômato cria submáquinas e as transições correspondentes. As alternativas de configuração são apresentadas na Fig.6.3, à direita. As submáquinas M4, M5, M6 e M7 reconhecem os tokens através de outro tipo de autômato que processa os tokens byte a byte (Fig. 6.4).

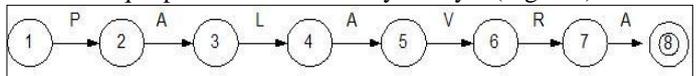


Figura 6.4. Reconhecedor de Palavras Simples.

No exemplo apresentado na Fig.6.4, o token “Palavra” é processado pela máquina M4; se o processamento terminar em um estado de aceitação, o token é reconhecido. As submáquinas M4, M5, M6 e M7 são criadas previamente por um programa que lê o Corpus e o converte em autômatos finitos determinísticos. A estrutura de identificação morfológica é composta por um par [chave, valor], no qual a chave é o estado de aceitação do elemento lexical e, o valor, sua classificação morfológica. No exemplo apresentado, a chave do item lexical “palavra” seria o estado “8”, e, o valor, a classificação morfológica, H+n (substantivo, núcleo de sintagma nominal), proveniente do Corpus Bosque.

O Formador de Palavras, submáquina M2, também é montado previamente por um programa construtor, levando em consideração as regras de formação de palavras do Português do Brasil, descritas por Margarida Basilio em [19]. Segundo a autora, o léxico é constituído por uma lista de formas já feitas e por um conjunto de padrões, os processos de formação de palavras, que determinam estruturas e funções tanto das formas já existentes quanto de formas ainda a serem construídas.

A submáquina M2 utiliza o vocabulário do TeP2.0 (formado por substantivos, adjetivos e verbos), como léxico de formas já feitas e o conjunto de regras de prefixação, sufixação e regressão verbal descrito pela autora para construir novas formas. No entanto, são necessários alguns cuidados para evitar a criação de estruturas que aceitem palavras inexistentes. A Fig. 6.5 apresenta um exemplo de autômato no qual são aplicados os prefixos “a” e “per”, e os sufixos “ecer” e “ar” (derivação parassintética) ao radical “noit” do substantivo noite, que faz parte do TeP2.0.

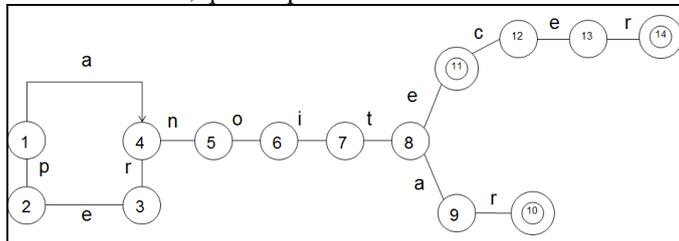


Figura 6.5. Autômato Formador de Palavras.

No exemplo apresentado, as palavras “anoitecer”, “pernoitar” e “anoitar” (sinônimo de “anoitecer”) existem no léxico do português do Brasil. Já pernoitecer é uma combinação que não existe na língua portuguesa. Margarida Basílio diz que algumas combinações não são aceitas simplesmente porque já existem outras construções consagradas pelo uso [20]. No entanto, acrescenta, existem centenas de substantivos terminados em -ção que não admitem formação adjetiva com o sufixo -oso. Por exemplo, vocação/vocacioso; intenção/intencioso; atração/atracioso. Segundo a autora, as derivações mais prováveis são aquelas em que há generalidade das funções envolvidas no processo de formação. Noções como a negação, o grau e a nominalização de verbos são bastante comuns e de grande generalidade, consequentemente, mais prováveis de serem válidas.

Para reduzir o risco de aceitar derivações inexistentes, o processo de construção do autômato restringe as possíveis formações, utilizando apenas as regras que a autora destaca como sendo mais prováveis. É o caso de nominalização de verbos com o uso dos sufixos -ção, -mento e -da. Em [19], Basílio cita que o sufixo -ção é responsável por 60% das formações regulares, enquanto o sufixo -mento é responsável por 20% destas formações. Já o sufixo -da é, via de regra, usado em nominalizações de verbos de movimento, tais como, entrada, saída, partida, vinda, etc.

Outra característica do construtor do autômato é representar os prefixos e sufixos sempre pelo mesmo conjunto de estados e transições, evitando repetições que acarretariam o consumo desnecessário de recursos computacionais. A Fig. 6.6 apresenta exemplo de palavras formadas por reutilização de estados e transições na derivação das palavras rueira, novidadeira e noveleira. Os estados e transições usados para

representar o sufixo “eira”, usado para designar são os mesmos nas três derivações.

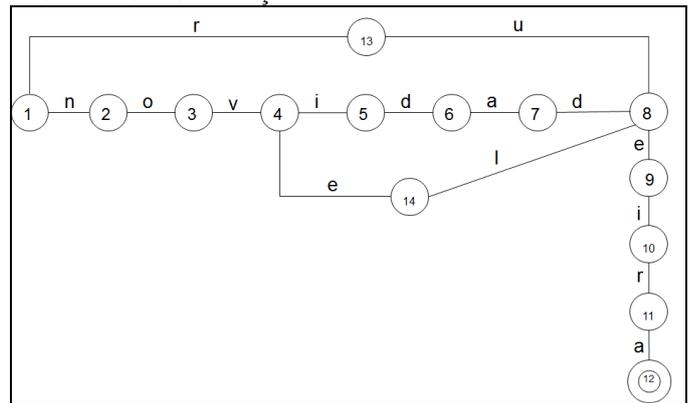


Figura 6.6. Autômato Formador de Palavras.

A submáquina M2 também reconhece palavras flexionadas, obtidas, no caso de substantivos e adjetivos, através da aplicação de sufixos indicativos de gênero, número e grau aos radicais do vocabulário TeP2.0. A Fig. 6.7 apresenta um exemplo de autômato usado para a formação das formas flexionadas do substantivo menino. Foram adicionados ao radical “menin” as flexões “o(s)”, “a(s)”, “ão”, “ões”, “onona(s)”, “inho(s)” e “inha(s)”.

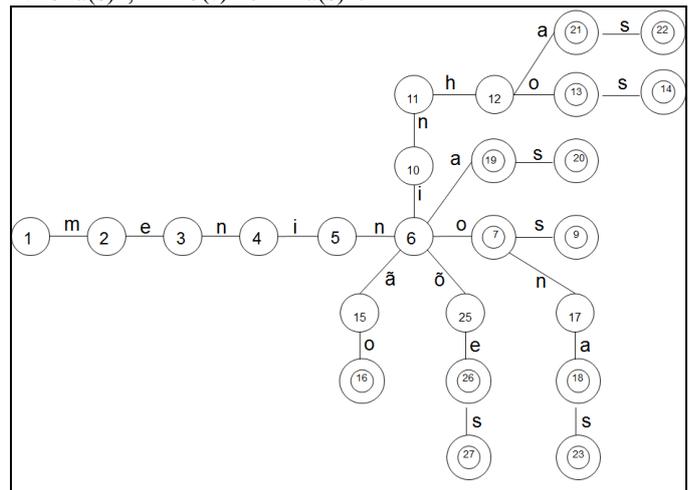


Figura 6.7. Autômato Formador de Flexões Nominais.

No caso de verbos, foi criada uma estrutura de estados e transições para representar as flexões de tempo, modo, voz e pessoa, obtendo-se, assim, as respectivas conjugações.

A Fig. 6.8 apresenta um exemplo de autômato usado para a formação das formas flexionadas do presente do indicativo do verbo andar. Foram adicionados ao radical “and” as flexões “o”, “as”, “a”, “andamos”, “ais” e “andam”.

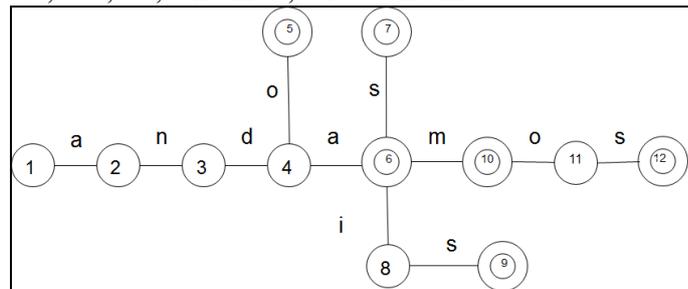


Figura 6.8. Autômato Formador de Flexões Verbais.

A estrutura de armazenamento da submáquina M2 também é composta por um par [chave, valor], no qual a chave é o

estado de aceitação do elemento lexical e, o valor, sua classificação morfológica, acrescida da origem do termo, indicando que ele foi gerado pelo construtor. Por exemplo, o termo “novidadeira” seria classificado como H+n+C – substantivo, núcleo de sintagma nominal, gerado pelo construtor.

Já a submáquina M3 é um autômato que varia em função do tipo de termo (conjunções, preposições e pronomes) e utiliza uma estrutura arbórea similar a M4. A Fig. 6.9 apresenta um exemplo de autômato usado no reconhecimento de termos deste domínio.

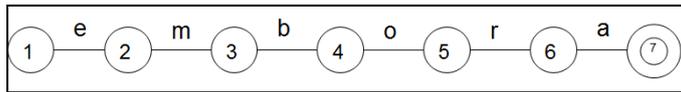


Figura 6.9. Autômato Reconhecedor de Conjunções, Preposições e Pronomes.

A estrutura de armazenamento da submáquina M3 é composta por um par [chave, valor], no qual a chave é o estado de aceitação do elemento lexical e, o valor, sua classificação morfológica, acrescida da origem do termo, indicando que ele faz parte da base de dados de apoio do Portal São Francisco. Por exemplo, o termo “embora” seria classificado como cj+Ba –conjunção da base de dados de apoio.

DESAMBIGUAÇÃO MORFOLÓGICA – ALGORITMO VITERBI

O Autômato Mestre também é responsável por desambiguar as classificações morfológicas e informar ao Agrupador apenas as que forem mais adequadas. Como as palavras podem ter mais do que uma classificação morfológica, é necessário utilizar uma técnica para selecionar aquela que é mais apropriada para o contexto analisado. Por exemplo, a palavra “casa” pode ser classificada como substantivo comum, feminino, singular ou como verbo flexionado na 3ª pessoa do singular no tempo presente, modo indicativo. No entanto, tendo em vista o contexto em que palavra se encontra, é possível escolher a classificação mais provável. Por exemplo, se a palavra “casa” vier precedida de um artigo definido, é mais provável que ela seja um substantivo; já se a palavra antecessora for um substantivo próprio, é mais provável que “casa” seja um verbo.

Collins [21] apresenta um modelo estatístico que leva em consideração 2 classificações anteriores à palavra analisada para fazer a desambiguação, criando distribuições nas quais a etiqueta de máxima probabilidade é o resultado da função:

- P = max p(E,S), sendo:
- E = etiquetas
- S = palavras da sentença
- p = probabilidade de E, dado S
- max = máxima probabilidade

Por exemplo, para etiquetar a frase “O advogado entrevista a testemunha”, a função receberia as palavras da sentença e as sequências de etiquetas possíveis para elas (obtidas a partir de um Corpus de testes), calculando, então, a sequência de maior probabilidade. No entanto, Collins alerta que a complexidade para executar tal função inviabiliza sua utilização, pois ela cresce exponencialmente em função do número de palavras da sentença (Em uma sentença de “n” palavras e “K” possíveis classificações, existiriam |K|^n possíveis etiquetas de sequência).

Para evitar o problema da complexidade da execução, Collins

propõe o uso do algoritmo Viterbi [22]. O algoritmo Viterbi é usado para encontrar a sequência mais provável de estados ocultos que resultam da observação de uma sequência de eventos. No caso da identificação das etiquetas morfológicas, os eventos são as palavras do texto analisado e os estados são as etiquetas de identificação morfológica. O algoritmo Viterbi pode ser implementado por meio de um autômato finito determinístico. No entanto, um autômato com estas características poderia ficar muito grande e, conseqüentemente, lento, devido ao tamanho do léxico usado para montá-lo. Uma alternativa para evitar este tipo de problema, é usar a tecnologia adaptativa. A Fig. 6.10 apresenta a estrutura do autômato que implementa o algoritmo Viterbi usado pelo desambiguador morfológico do Linguístico.

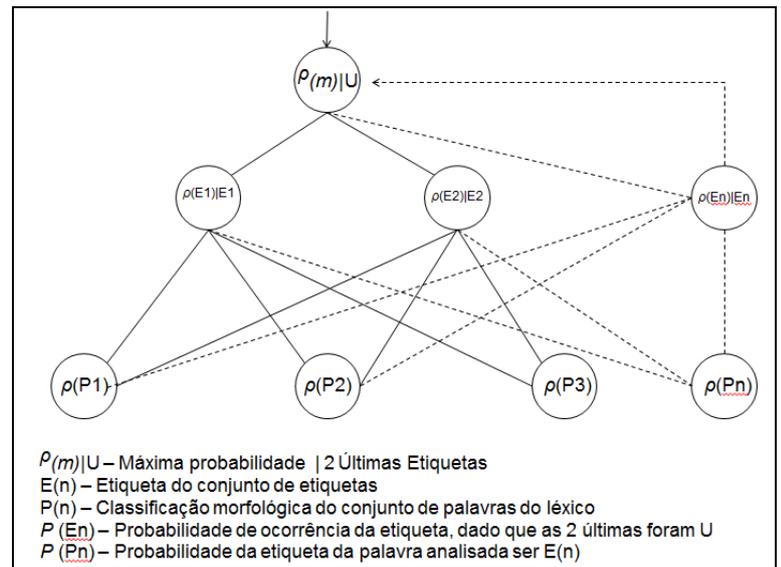


Figura 6.10. Autômato Adaptativo Viterbi.

O autômato proposto recebe como parâmetro de entrada a palavra que está sendo analisada e monta dinamicamente os estados e transições de acordo com as possíveis etiquetas e as respectivas probabilidades. Por exemplo, no caso da frase citada anteriormente “O advogado entrevista a testemunha”, o autômato se comportaria da seguinte forma:

- Primeiro token: “O”
- $\rho(m)|U = 1$ | ** (* * indica inicio de frase)
- E1 = Artigo
- E2 = Pronome
- E3 = Preposição
- $\rho(E1) = 0,55$: probabilidade de artigo seguido de * *
- $\rho(E2) = 0,25$: probabilidade de pronome seguido de * *
- $\rho(E3) = 0,20$: probabilidade preposição seguido de * *
- P1 – O (artigo)
- P2 – O (pronome)
- P3 – O (preposição)
- $\rho(P1) = 0,60$: probabilidade de “O” ser artigo
- $\rho(P2) = 0,30$: probabilidade de “O” ser pronome
- $\rho(P3) = 0,10$: probabilidade de “O” ser preposição
- $\rho(m) = 0,55*0,6 | 0,25*0,3 | 0,2*0,1$
- $\rho(m) = 0,55*0,6 = 0,33$ | U = Artigo

- Segundo token: “advogado”

$\rho(m)|U = 0,33 | * \text{ Artigo}$

E1 = Substantivo

E2 = Adjetivo

E3 = Advérbio

$\rho(E1) = 0,60$: probabilidade de substantivo seguido de * e artigo

$\rho(E2) = 0,20$: probabilidade de adjetivo seguido de * e artigo

$\rho(E3) = 0,20$: probabilidade de advérbio seguido de * e artigo

P1 – Advogado (substantivo)

$\rho(P1) = 1,0$: probabilidade de “Advogado” ser substantivo

$\rho(P2) = 0$: probabilidade de “Advogado” ser adjetivo

$\rho(P3) = 0$: probabilidade de “Advogado” ser advérbio

$\rho(m) = 0,33*0,60*1 | 0,33*0,20*0 | 0,33*0,20*0$

$\rho(m) = 0,33*0,60*1 = 0,198 | U = \text{ Artigo Substantivo}$

- Terceiro token: “entrevista”

$\rho(m)|U = 0,198 | \text{ Artigo Substantivo}$

E1 = Substantivo

E2 = Verbo

E3 = Adjetivo

$\rho(E1) = 0,60$: probabilidade de verbo seguido de artigo e substantivo

$\rho(E2) = 0,20$: probabilidade de adjetivo seguido de artigo e substantivo

$\rho(E3) = 0,20$: probabilidade de adjetivo seguido de artigo e substantivo

P1 – entrevista (verbo)

P2 – entrevista (substantivo)

$\rho(P1) = 0,70$: probabilidade de “entrevista” ser verbo

$\rho(P1) = 0,30$: probabilidade de “entrevista” ser substantivo

$\rho(P1) = 0$: probabilidade de “testemunha” ser adjetivo

$\rho(m) = 0,198*0,60*0,70 | 0,198*0,20*0,30 | 0,198*0,20*0$

$\rho(m) = 0,198*0,60*0,70 = 0,0831 | U = \text{ Substantivo Verbo}$

- Quarto token: “a”

$\rho(m)|U = 0,0831 | \text{ Substantivo Verbo}$

E1 = Artigo

E2 = Substantivo

E3 = Adjetivo

$\rho(E1) = 0,60$: probabilidade de artigo seguido de substantivo e verbo

$\rho(E2) = 0,20$: probabilidade de substantivo seguido de substantivo e verbo

$\rho(E3) = 0,20$: probabilidade de adjetivo seguido de substantivo e verbo

P1 – A (artigo)

P2 – A (pronomes)

P3 – A (preposição)

$\rho(P1) = 0,60$: probabilidade de “a” ser artigo

$\rho(P2) = 0,30$: probabilidade de “a” ser pronomes

$\rho(P3) = 0,10$: probabilidade de “a” ser preposição

$\rho(m) = 0,0831*0,60*0,60 | 0,0831*0,20*0,30 | 0,0831*0,20*0,10$

$\rho(m) = 0,0831*0,60*0,60 = 0,0299 | U = \text{ Verbo Artigo}$

- Quinto token: “testemunha”

$\rho(m)|U = 0,0299 | \text{ Verbo Artigo}$

E1 = Substantivo

E2 = Verbo

E3 = Adjetivo

$\rho(E1) = 0,60$: probabilidade de substantivo seguido de verbo e artigo

$\rho(E2) = 0,20$: probabilidade de verbo seguido de verbo e artigo

$\rho(E3) = 0,20$: probabilidade de adjetivo seguido de verbo e artigo

P1 – testemunha (verbo)

P2 – testemunha (substantivo)

$\rho(P1) = 0,70$: probabilidade de “testemunha” ser substantivo

$\rho(P2) = 0,30$: probabilidade de “testemunha” ser verbo

$\rho(P3) = 0$: probabilidade de “testemunha” ser adjetivo

$\rho(m) = 0,0299*0,60*0,70 | 0,0299*0,20*0,30 | 0,0299*0,20*0$

$\rho(m) = 0,0299*0,60*0,70 = 0,0125 | U = \text{ Artigo Substantivo}$

Portanto, ao final do processamento, a sequência de etiquetas seria: Artigo Substantivo Verbo Artigo Substantivo.

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.

O Linguístico encontra-se em fase de construção, dividida em etapas em função da estrutura do reconhecedor. A primeira versão, englobando sentenciador, tokenizador e identificador morfológico, foi finalizada.

O trabalho encontra-se na fase de projeto do Desambiguador Morfológico, que foi totalmente concebido para utilizar tecnologia adaptativa.

REFERÊNCIAS

- [1] NETO, J.J. APRESENTAÇÃO LTA-LABORATÓRIO DE LINGUAGENS E TÉCNICAS ADAPTATIVAS. DISPONÍVEL EM: [HTTP://WWW.PCS.USP.BR/~LTA](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 Nº 38, SÃO CARLOS, 88P, 1999. 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] LINGUATECA : [HTTP://WWW.LINGUATECA.PT/](http://www.linguateca.pt/)
- [14] TEP2. THESAURO ELETRÔNICO PARA O PORTUGUÊS DO BRASIL. DISPONÍVEL EM: <[HTTP://WWW.NILC.ICMC.USP.BR/TEP2/](http://www.nilc.icmc.usp.br/tep2/)>
- [15] PORTAL SÃO FRANCISCO. MATERIAIS DE LÍNGUA PORTUGUESA. DISPONÍVEL EM: <[HTTP://WWW.PORTALSAOFRANCISCO.COM.BR/ALFA/MATERIAS/INDEX-LINGUA-PORTUGUESA.PHP/](http://www.portalsaofrancisco.com.br/alfa/materias/index-lingua-portuguesa.php/)>
- [16] WIKIPEDIA. DISPONÍVEL EM: <[HTTP://PT.WIKIPEDIA.ORG/WIKI/P%C3%A1gina_principal/](http://pt.wikipedia.org/wiki/P%C3%A1gina_principal/)>
- [17] KARLSSON, F. CONSTRAINT GRAMMAR AS A FRAMEWORK FOR PARSING RUNNING TEXT. 13o. INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, HELSINKI (VOL.3, PP.168-173).
- [18] PYTHON. PYTHON PROGRAMMING LANGUAGE OFFICIAL WEB SITE. DISPONÍVEL EM: <<http://www.python.org/>>
- [19] BASILIO, M. FORMAÇÃO E CLASSES DE PALAVRAS NO PORTUGUÊS DO BRASIL. Ed.CONTEXTO, 2004
- [20] BASILIO, M. TEORIA LEXICAL. Ed.ÁTICA, 1987
- [21] COLLINS, M. COURSE NOTES FOR NLP. COLUMBIA UNIVERSITY. DISPONÍVEL EM: [HTTP://WWW.CS.COLUMBIA.EDU/~MCOLLINS/HMMS-SPRING2013.PDF](http://www.cs.columbia.edu/~mcollins/hmms-spring2013.pdf)
- [22] FORNEY, G.D., JR. IEEE. PROCEEDINGS OF THE IEEE (VOLUME:61 , ISSUE: 3), 1973

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 Experían.

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.