SOFIA: UM IDENTIFICADOR DE NOMES E VERBOS PARA O PORTUGUÊS DO BRASIL

Gustavo Paiva Guedes e Silva

Dissertação de Mestrado submetida ao Programa de Pós-Graduação Lingüística da Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários para a obtenção do título de Mestre em Lingüística.

Rio de Janeiro
Dezembro de 2008

SOFIA: Um identificador de nomes e verbos para o Português do Brasil

Gustavo Paiva Guedes e Silva
Orientadora: Professora Doutora Lilian Vieira Ferrari

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Lingüística da Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários para a obtenção do título de Mestre em Lingüística.

Examinada por:

_____
Presidente, Profa. Lilian Vieira Ferrari

_____
Prof. Carlos Alexandre V. Gonçalves – UFRJ

_____
Prof. Luiz Fernando M. Rocha – UFJF

_____
Profa. Maria Lúcia Leitão de Almeida – UFRJ, Suplente

_____
Prof. Celso Vieira Novaes – UFRJ, Suplente

Rio de Janeiro
Dezembro de 2008

Dedicatória

Dedico aos meus pais, por tudo.

Agradecimentos

Agradeço a todos os meus amigos que contribuíram de forma direta e indireta para a construção deste trabalho.

Agradeço aos professores que como verdadeiros mestres me transmitiram seus conhecimentos.

Agradeço à Professora Dra. Lilian Vieira Ferrari pelo ensinamento, apoio, orientação, críticas e sugestões.

Agradeço a Leandro Diniz que bastante me ensinou sobre informática e que me auxiliou na construção do banco de dados lexical.

“Aqueles que imaginam que todos os frutos
amadurecem ao mesmo tempo como as cerejas,
é porque nada sabem acerca das uvas.”

Paracelso

Abstract

This work describes the development of a computational system for the identification of nouns and verbs in Brazilian Portuguese written texts. The system comprises four modules (controller, lexical analysis, morphological analysis and syntactic analysis), which interact to perform the correct identification of both unproblematic and problematic cases related to categorial ambiguity.

Resumo

Esta dissertação de tese descreve a criação de um sistema para identificação de nomes e verbos em textos escritos do Português do Brasil. O sistema se divide em quatro módulos principais: controlador, módulo de análise léxica, análise morfológica e análise sintática. Esses módulos interagem para proceder à identificação das categorias em questão, permitindo a identificação correta tanto de casos não problemáticos quanto de casos relacionados à ambigüidade categorial.

Palavras Chave

Processamento de linguagem natural; texto escrito; analise léxica; analise sintática; análise morfológica; morfologia; sintaxe; Gramática das Construções; nomes; verbos.

SUMÁRIO

1. INTRODUÇÃO 15

2. FUNDAMENTOS LINGÜÍSTICOS 17

2.1. HOMONÍMIA E POLISSEMIA 17

2.1.1 Alguns problemas de classificação 19

2.2. GRAMÁTICA DAS CONSTRUÇÕES 21

2.2.1 Trabalhos anteriores 23

3. PROCESSAMENTO DE LINGUAGEM NATURAL 28

3.1. HISTÓRIA 28

3.2. DESCRIÇÃO 29

4. O SISTEMA PROPOSTO 32

4.1. INTRODUÇÃO 32

4.2. A ELABORAÇÃO DO MÉTODO 33

4.3. ARQUITETURA E FUNCIONAMENTO DO SISTEMA PROPOSTO 37

4.4. ELABORAÇÃO DE BASE DE DADOS LEXICAIS 42

4.5. METODOLOGIA 47

4.5.1 Artigo 50

4.5.2 Substantivo 51

4.5.3 Adjetivo 53

4.5.4 Pronome 53

4.5.5 Outras classes gramaticais 54

4.6. MÓDULO GERENCIADOR 56

4.7. MÓDULO DE MEMÓRIA RECENTE 57

4.8. MÓDULO DE ANÁLISE LÉXICA 58

4.9. MÓDULO DE ANÁLISE MORFOLÓGICA 59

4.9.1 Número dos substantivos e adjetivos 60

4.9.2 Gênero dos substantivos e adjetivos 64

4.9.3 Grau dos substantivos e adjetivos 67

4.10. MÓDULO DE ANÁLISE SINTÁTICA 69

5. RESULTADOS OBTIDOS 75

5.1. TEXTO JORNALÍSTICO 75

5.1.1 Problemas de classificação 78

5.2. TEXTO LITERÁRIO 86

5.2.1 Problemas de classificação 88

6. CONCLUSÃO 94

7. REFERÊNCIAS BIBLIOGRÁFICAS 96

APÊNDICE A – RÓTULOS UTILIZADOS NO SISTEMA 98

APÊNDICE B – NEM A ROSA NEM O CRAVO... (JORGE AMADO) 100

APÊNDICE C – CETEM FOLHA 102

LISTA DE ABREVIATURAS

BD	Banco de Dados
COPPE	Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
GC	Gramática das Construções
IPA	*International Phonetic Alphabet*
MIT	*Massachusetts Institute of Technology*
PB	Português do Brasil
PLN	Processamento de Linguagem Natural
UFRJ	Universidade Federal do Rio de Janeiro
WWW	*World Wide Web*
XML	*Extensible Markup Language*

LISTA DE FIGURAS

FIGURA 4.1 – EXEMPLO DE AMBIGÜIDADE CATEGORIAL. 33

FIGURA 4.2 – EXEMPLO DE CLASSE E OBJETO 36

FIGURA 4.3 – ARQUITETURA CLIENTE/SERVIDOR 38

FIGURA 4.4 – ARQUITETURA MACRO DO SISTEMA 39

FIGURA 4.5 – ARQUITETURA INTERNA DO SISTEMA 40

FIGURA 4.6 – RESULTADO OBTIDO PELO SISTEMA 41

FIGURA 4.7 – EXEMPLO DE BUSCA EFETUADA PELO SISTEMA. 43

FIGURA 4.8 – ARQUIVO XML DE ABREVIAÇÕES 46

FIGURA 4.9 – CONSULTA DO VOCÁBULO “AUMENTE” 55

FIGURA 4.10 – ARQUIVO XML DE PREPOSIÇÕES 57

FIGURA 4.11 – ARQUITETURA DO SISTEMA 59

FIGURA 4.12 – ARQUIVO XML DE PLURAIS 61

FIGURA 4.13 – ARQUITETURA DO SISTEMA 62

FIGURA 4.14 – PROCESSAMENTO DO GÊNERO 65

FIGURA 4.15 – MÓDULO DE ANÁLISE SINTÁTICA 69

FIGURA 4.16 – ALGORITMO DE EXEMPLO 70

FIGURA 4.17 – RESULTADO DE BUSCA PARA OS VOCÁBULOS “VELHO” E “MUNDO” 72

FIGURA 4.18 – USO DE REGRAS EM NOMES AMBÍGUOS 85

FIGURA 4.19 – USO DE REGRAS EM VERBOS AMBÍGUOS 86

LISTA DE TABELAS

TABELA 4.1 – CLASSES DE PALAVRAS UTILIZADAS 48

TABELA 4.2 – RÓTULOS DE ENTRADAS NO BD 49

TABELA 4.3 – EXEMPLO DE ENTRADA NO BD 50

TABELA 4.4 – PRONOMES POSSESSIVOS 54

TABELA 4.5 – EXEMPLO DE ALGUMAS REGRAS DE PLURAL 63

TABELA 4.6 – TRANSFORMAÇÃO DE GÊNERO FEMININO PARA MASCULINO 66

TABELA 4.7 – EXEMPLO DE REGRA DE TRANSFORMAÇÃO NÃO CADASTRADA 66

TABELA 4.8 – EXEMPLO DE RESULTADO DE TRANSFORMAÇÃO PARA “ALUNA” 67

TABELA 4.9 – EXEMPLO DE REGRAS DE GRAU 68

TABELA 4.10 – RESULTADOS OBTIDOS 76

TABELA 4.11 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MAIS 78

TABELA 4.12 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MENOS 81

TABELA 4.13 – RESUMO DA CLASSIFICAÇÃO DE VERBOS A MAIS 82

TABELA 4.14 – RESUMO DA CLASSIFICAÇÃO DE VERBOS A MENOS 84

TABELA 4.15 – RESULTADOS OBTIDOS NO TEXTO LITERÁRIO 87

TABELA 4.16 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MAIS 88

TABELA 4.17 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MENOS 90

TABELA 4.18 – RESUMO DA CLASSIFICAÇÃO DE VERBOS A MAIS 91

LISTA DE ESQUEMAS

ESQUEMA 2.1 – REGRA UTILIZADA NO TRABALHO DE 2004. 24

ESQUEMA 2.2 – REGRA UTILIZADA NO TRABALHO DE IC. 25

ESQUEMA 2.3 – ESQUEMA PARA <COMEÇO>. 25

1. INTRODUÇÃO

Esta dissertação representa uma proposta de associação entre os campos teóricos da "Lingüística" e da "Ciência da Computação", com a finalidade de contribuir para o desenvolvimento de novas tecnologias.

Em especial, a elaboração de tradutores automáticos e de sistemas de conversão texto-fala representam importantes produtos dessa cooperação. Embora já existam ferramentas desenvolvidas nessas áreas disponíveis no mercado, o que se observa é que os resultados alcançados ainda são pouco satisfatórios.

Dentre os problemas apresentados, destaca-se a incapacidade de lidar com palavras homônimas ou polissêmicas. Por exemplo, se solicitarmos a tradução de "O ator casa no sábado" à ferramenta de tradução do Google™, obteremos "The actor home on saturday". Isso evidencia que o tradutor só está preparado para buscar um dos significados da palavra "casa", desconsiderando a homonímia com a 3ª pessoa do singular do verbo "casar".

O primeiro passo a ser dado seria capacitar o sistema a identificar classes de palavras. Assim, a identificação de "casa" como "verbo" no exemplo supracitado ("o ator casa no sábado") permitiria a tradução correta.

Do mesmo modo, em sistemas de conversão texto-fala, a incapacidade de distinguir entre diferentes pronúncias de homônimos homógrafos, como "começo" ou "gosto" determinará conversões inadequadas, pois esses homônimos são foneticamente distintos. Para que a questão possa ser resolvida, mais uma vez, faz-se necessário identificar se a ocorrência das palavras envolvidas se inclui na classe dos "nomes" ou dos "verbos".

Constatados os problemas acima, o objetivo deste trabalho é descrever os passos e metodologias utilizadas para a construção do sistema SOFIA[1] (http://www.projetosofia.com.br), capaz de identificar nomes e verbos em sentenças do Português do Brasil. A metodologia proposta utilizará as bases teóricas da Gramática das Construções (GC), associada à Lingüística Cognitiva e o Processamento de Linguagem Natural (PLN), que estuda a compreensão e geração automática de linguagem natural.

O sistema proposto não utiliza métodos baseados em estatísticas, pois tem o propósito de construir regras gramaticais que possam dar conta de entradas textuais que não sejam restritas a um contexto específico.

A dissertação está organizada da seguinte forma: no capítulo 2, serão apresentados os fundamentos lingüísticos que servirão de base para a criação do sistema proposto, discutindo-se, em especial, a problemática dos pares homônimos, os fundamentos teóricos da vertente da GC e trabalhos anteriores sobre desambiguação de homônimos em sistemas de conversão texto-fala. O capítulo 3 apresenta uma descrição sobre o processamento de linguagem natural. No capítulo 4, o sistema é detalhado, especificando-se a arquitetura utilizada em termos de seus módulos componentes (léxico, morfológico e sintático). Por fim, o capítulo 5 apresenta a análise dos dados, destacando os resultados obtidos pelo sistema com base em dois textos (jornalístico e literário). Nesse capítulo, são discutidos também os problemas de análise.

---

[1] A escolha do nome SOFIA foi inspirada na palavra grega SOPHIA que significa sabedoria. Foi escolhido para destacar que o sistema deve ser provido de determinado conhecimento/sabedoria para ser capaz de discernir entre as diversas nuances das linguagens naturais, no caso o PB.

2. FUNDAMENTOS LINGÜÍSTICOS

Neste capítulo, serão apresentados os fundamentos teóricos para o tratamento lingüístico dos dados. Na seção 2.1, serão tratados, sucintamente, os fenômenos semânticos da homonímia e da polissemia. Na seção 2.2, será apresentado o paradigma teórico da Gramática das Construções.

2.1. Homonímia e Polissemia

Segundo Basílio (2004), quando os significados de uma mesma forma fonológica não são relacionados, denominamos a situação como homonímia. Caso os significados sejam relacionados, é preferível considerarmos um caso onde uma única palavra possui vários sentidos, ou seja, damos à situação o nome de polissemia. A autora apresenta os exemplos "**regra** da gramática normativa" e "**regra** de etiqueta" como uma situação em que a palavra "regra" é polissêmica, visto que, em ambos os casos, existe um significado geral de prescrição, distintos apenas pelo domínio ao qual se aplica. Para o caso de homonímia, Basílio nos apresenta o exemplo "manga", que pode ser fruta ou parte do vestuário. Ao final do tópico no qual discute a homonímia e a polissemia, a autora afirma que a distinção continua sendo discutida, tanto teoricamente como em casos particulares, e que apresenta um problema permanente em relação ao conceito de *palavra*. Podemos subdividir os homônimos em três casos: os casos em que palavras de sentidos distintos possuem a mesma grafia (os homônimos homógrafos) ou a mesma pronúncia (os homônimos homófonos). Há também os

chamados homônimos perfeitos, que designam as palavras que possuem pronúncias iguais e mesma grafia. Abaixo podemos ver exemplos dos homônimos[2]:

1) homônimos homógrafos (heterófonos):

- gosto (substantivo)
- gosto (1.a pessoa do singular do presente do indicativo – verbo gostar)

2) homônimos homófonos (heterógrafos):

- cela (substantivo)
- sela (1ª pessoa do singular do presente do indicativo – verbo selar)

3) homônimos homófonos e homógrafos (homônimos perfeitos):

- verão (substantivo)
- verão (3ª pessoa do plural do presente do indicativo – verbo ver)

Como o sistema desenvolvido utiliza entradas textuais (casos 1 e 3), os homônimos homógrafos precisarão de tratamento especial. Ao analisarmos as duas frases abaixo, notamos que o vocábulo "contorno" se insere no caso 1:

---

[2] Embora haja uma vasta literatura em que se discutem as possíveis diferenças entre homonímia e polissemia, não é objetivo deste trabalho tratar do assunto, já que a idéia é apenas enfatizar a existência de vocábulos homógrafos que pertencem a classes gramaticais distintas.

(1) "...determinar o perímetro do ***contorno*** da região hachurada..." [3]

(2) "Não entendi como eu ***contorno*** isso..." [4]

O fato dos homônimos homógrafos possuírem pronúncia diferente será irrelevante para o sistema proposto nesta dissertação, visto que o sistema trabalha apenas com entradas textuais. Portanto, vocábulos como cont[o]rno x cont[ɔ]rno devem ser desambigüizados com base em regras de decisão que levam em conta apenas as distinções existentes a partir de informações textuais.

O sistema não apresentará qualquer dificuldade perante os homônimos heterógrafos, visto que para um sistema que processa informações textuais, a diferença na grafia resultará na correta classificação gramatical.

2.1.1 Alguns problemas de classificação

As classes de palavras não formam categorias estanques. Na verdade, a língua apresenta paradigmas lexicais nos quais uma mesma raiz pode servir de base para palavras de classes gramaticais diferentes, por meio de mecanismos morfológicos de prefixação, sufixação, alternâncias vocálicas, etc. Além disso, há também casos como os discutidos na seção anterior, em que um mesmo item lexical, com sua forma externa inalterada, pode pertencer a uma ou outra

---

[3] Referência encontrada no sistema de busca YAHOO!®

[4] Referência encontrada no sistema de busca YAHOO!®

classe gramatical. Entre os casos fronteiriços discutidos na literatura encontram-se "nome e adjetivo" e "verbo e adjetivo".

Com relação à distinção entre *nome* e *adjetivo*, Lemle (1984:102) estabelece a seguinte regra de nominalização de adjetivos:

> *Todo adjetivo que pode modificar um nome referente a um ser humano pode exercer papel de nome, incorporando o conceito de pessoa ao seu próprio sentido, que passa a ser: uma pessoa com a qualidade expressa pelo adjetivo.*

A regra acima aponta a possibilidade de que o adjetivo em questão funcione como núcleo do sintagma nominal e admita um determinante como especificador. Com relação ao SOFIA, essa possibilidade foi contemplada.

O problema, entretanto, se mostrou mais resistente em um número restrito de situações, que serão objeto de análise no item 4.10 (Módulo de Análise Sintática).

Com relação à distinção entre a classe dos verbos e a dos adjetivos, o caso fronteiriço é o dos particípios passados. A teoria lexicalista considera que tantos os particípios em posição adnominal (ex. riscos calculados) quanto predicativos do sujeito em passivas (ex. Os riscos foram calculados) são adjetivos. (Lemle, 1984; Pimenta-Bueno, 1981).

Apesar da decisão adotada pelos autores mencionados acima, mantivemos a distinção entre adjetivo (particípios adnominais) e verbos (particípios em passivas) nessa primeira etapa do trabalho. A motivação para essa decisão foi a manutenção da flexibilidade do sistema para futuras aplicações, como por exemplo, a tradução automática para o inglês. Se quisermos traduzir "O filme foi exibido", o particípio "exhibited" será adequado; entretanto, um caso como "a menina exibida" não poderia ser traduzido por *"an exhibited girl". Para prevenir erros desse tipo, optou-se por

manter os particípios em posição adnominal nas classes dos adjetivos e os particípios que integram as passivas, na classe dos verbos. Essa decisão, contudo, acabou gerando erros de classificação que também serão descritos no item 4.10.

2.2. Gramática das Construções

O ponto de vista adotado pela Lingüística Cognitiva é o de que o conhecimento que o falante tem de uma língua é caracterizado como um inventário estruturado de unidades convencionais [Langacker, 1987]. Tais unidades incluem morfemas, palavras, sintagmas e esquemas genéricos que descrevem os padrões gramaticais convencionais, utilizados, inclusive, para a criação de sentenças e novos sintagmas. Os esquemas são adquiridos através da exposição às expressões reais que os instanciam.

O paradigma teórico da "Gramática das Construções" postula que as unidades apropriadas da gramática são *construções gramaticais*, que especificam não apenas informação sintática, mas também informação lexical, semântica e pragmática.

As expressões compostas são incluídas na gramática de uma língua, na proporção em que adquirem status de unidades convencionais. As regularidades na formação das expressões compostas são representadas na gramática por hierarquias de construções esquemáticas, caracterizadas em níveis apropriados de abstração; tanto os subesquemas quanto as expressões específicas podem instanciar um esquema particular. Por exemplo, a caracterização mais esquemática do sintagma preposicional em Português especifica, simplesmente, a seqüência <P + N> (ou seja, preposição seguida de nome). Entretanto, vários subesquemas poderiam ser

reconhecidos, tais como <P + PRON>, <com +N>), ou mesmo <com + PRON>(que instancia os dois anteriores). As expressões específicas *com ele*, *com ela*, etc instanciam todos os subesquemas mencionados, tanto diretamente como através de relações elaborativas.

Tais exemplos indicam que as construções gramaticais são vistas como categorias complexas e representadas sob a forma de redes esquemáticas. O conhecimento do falante da construção de sintagma preposicional inclui não apenas um esquema de alto nível, mas também subesquemas, expressões específicas, e relações de categorização que associam todas essas estruturas.

Algumas teorias lingüísticas, como a gerativa, postulam a existência de fronteiras bem definidas entre léxico, sintaxe, semântica e pragmática. Outras, como a Gramática das Construções, postulam que não há separação entre o léxico e a sintaxe, assim como entre a semântica e a pragmática. Dessa forma, a Gramática das Construções apresenta a idéia de que todas as dimensões da língua contribuem diretamente para a descrição das expressões lingüísticas. Esse é, certamente, o principal motivo para a escolha dessa última teoria para o desenvolvimento desta dissertação. Como será demonstrado no decorrer do trabalho, a manutenção de fronteiras entre esses níveis de análise é o principal fator responsável pelo alto índice de erros gerados nos tradutores atuais. Casos como o verbo "tomar" são a prova mais concreta de que a semântica do verbo deve ser levada em consideração para traduções mais eficientes, assim como os casos "tomar coragem", "tomar banho", "tomar sorvete", "tomar liberdade", etc., que acabam incorretamente traduzidos por não ocorrer qualquer processamento semântico/pragmático nos tradutores existentes.

Embora o sistema proposto nesta dissertação não contemple as análises semântica e pragmática, estará preparado para essa possibilidade. A idéia é que futuramente esses módulos sejam integrados com o módulo de tradução.

2.2.1 Trabalhos anteriores

Alguns trabalhos para sistemas de conversão texto-fala já foram realizados para o Português usando o paradigma da Gramática das Construções. São eles: Barbosa, Ferrari, Resende (2004), e Guedes e Ferrari, (2004).

A motivação para a realização desses trabalhos surgiu através de pesquisas na área de sistemas texto/fala realizadas em parceria com a COPPE/UFRJ. Esses sistemas são responsáveis por transformar texto em fala e a fala em texto. A tarefa de transformar texto em fala apresentou alguns desafios, dentre os quais, o fato de existirem vocábulos no Português que são escritos com a mesma grafia, mas com diferenças na realização fonológica, ou seja, homógrafos e heterófonos, como é o caso de "gosto" (g[o]sto e g[ɔ]sto). Casos como esses são freqüentemente encontrados na língua portuguesa, e alguns deles formaram o escopo do trabalho realizado na iniciação científica.

O primeiro trabalho citado [Barbosa, Ferrari, Resende (2004)], realizado em parceria com a COPPE/UFRJ, iniciou com uma identificação de alguns pares de homógrafos-heterófonos, para assegurar que era um problema bastante encontrado no Português do Brasil. Com esse estudo inicial, foi possível evidenciar a ocorrência de dois grupos de homógrafos-heterófonos: 1. homógrafos pertencentes a classes gramaticais diferentes (ex: <começo>[5], cuja pronúncia pode ser com[ɛ]ço, indicando a 1ª pessoa do singular do presente do indicativo do verbo <começar> ou com[e]ço indicando o substantivo referente ao verbo em questão. b. homógrafos com uma dupla possibilidade de oposição, a primeira possibilidade contendo pares pertencentes a mesma classe

[5] Seqüências de grafemas serão envolvidas pelos símbolos '<>'.

gramatical e a segunda apresentando classes gramaticais diferentes (ex: c[o]rte (N) x c[ɔ]rte (N) ou c[o]rte (N) x c[ɔ]rte(V).

As regras construídas para resolver a ambigüidade basearam-se nos constituintes que precediam e nos constituintes que sucediam a palavra em questão, ou seja, toda vez que surgia a palavra <gosto> em um texto, o programa verificava algumas palavras que a precediam e algumas que a sucediam. O número de palavras que o sistema considerou para o estudo foi restringido pela profundidade inserida na regra a ser considerada. Como exemplo, podemos observar a regra abaixo:

(Vetor_Artigos, PROF_NIVEL_1, resultado1) <gosto>

ESQUEMA 2.1 – REGRA UTILIZADA NO TRABALHO DE 2004.

Conforme demonstra a figura acima, dentro de um vetor de palavras composto por artigos e com uma profundidade de nível 1 (pode variar de 1 até N) à esquerda (no caso em questão a regra está a esquerda do vocábulo em questão representado por "<gosto>"), o programa verifica se a palavra que antecede a seqüência de grafemas <gosto> é pertencente a este vetor de artigos; se for, a transcrição grafema-fone é feita para o 1º resultado , no caso, g[o]sto. Caso contrário, o programa continua verificando as regras subseqüentes. Foram encontradas 1438 ocorrências de <gosto>, sendo 849 de g[o]sto e 589 de g[ɔ]sto no 'Corpus de Textos Eletrônicos NILC/Folha de São Paulo(CETEN-Folha). O conjunto de regras desenvolvidas nesse trabalho obteve índice de acerto igual a 94.5%.

Em trabalho de Iniciação Científica desenvolvido entre os anos de 2003 e 2004, analisei os homógrafos-heterófonos c[o]rte x c[ɔ]rte e com[e]ço x com[ɛ]ço. A metodologia utilizada para o desenvolvimento desse projeto foi a mesma utilizada para o desenvolvimento do projeto que analisou os pares g[o]sto x g[ɔ]sto, apenas mudando a implementação do sistema e a nomenclatura das funções. Abaixo segue o exemplo da primeira regra utilizada no sistema, que enfoca a relação sujeito/verbo do vocábulo "começo" no nível da sentença, com base no esquema a seguir:

AcentuaEsq(pron_pes_1,1, [comεço])

ESQUEMA 2.2 – REGRA UTILIZADA NO TRABALHO DE IC.

A regra acima refere-se ao esquema abaixo:

PRON. PESSOAL(1ª PESSOA SING.)	VERBO
	↓
	[comεço]
S	

ESQUEMA 2.3 – ESQUEMA PARA <COMEÇO>.

Durante o processamento de uma frase, a regra ilustrada no esquema 2.2 determina que caso haja a presença do vocábulo "começo", o sistema irá verificar o vetor de palavras representado pela variável *pron_pes_1* com uma profundidade de um vocábulo a esquerda de "começo". Caso a palavra imediatamente precedente a "começo" esteja presente no vetor em questão, o sistema irá

convergir para o resultado com[ɛ]ço e não irá processar as regras restantes. Caso contrário, o sistema continua a processar as regras subsequentes. Caso nenhuma regra seja aplicável à estrutura, o sistema converge para um resultado padrão. Abaixo podemos ver dois exemplos de frases com o vocábulo "começo".

(3) Eu começo o trabalho em uma nova coreografia assim.

(4) O começo de uma outra humanidade.

Podemos ler sem qualquer confusão as duas frases acima. No exemplo 3, percebemos claramente a presença do verbo "começar" na 1ª pessoa do singular do presente do indicativo e no exemplo 4, subtítulo da obra de Michel Serres intitulada Hominescencias, não há qualquer dúvida sobre a presença do substantivo "começo". Fazemos esse processo de desambiguação categorial inconscientemente, mas como sabemos, um computador é incapaz de fazer essa distinção. Isso só se torna possível com a utilização de regras gramaticais.

Foram encontradas 1630 ocorrências de <corte> e 1449 de <começo> no 'Corpus de Textos Eletrônicos NILC/Folha de São Paulo(CETEN-Folha)[6]. É interessante notar que se o conversor texto-fala estivesse programado para escolher para cada situação apenas a forma com maior número de ocorrências, <corte> teria um acerto de 66% e <começo>, um acerto de 94,6%.

O estabelecimento de esquemas e a fixação de itens lexicais instanciados pelos mesmos, permitiu que o conversor texto-fala acertasse o membro correto do par de homógrafos c[o]rte e

[6] Corpus de Extractus de Textos Eletrônicos NILC/ Folha de São Paulo.

[cɔrte] em 98,5% das vezes. Resultados melhores foram obtidos no par [começo] e [comɛço], atingindo um índice de acerto de 99,86%.

3. PROCESSAMENTO DE LINGUAGEM NATURAL

Este capítulo apresenta uma descrição sobre o processamento de linguagem natural. Na seção 3.1 será apresentada uma breve descrição histórica sobre o surgimento do processamento de linguagem natural. Em seguida, a seção 3.2 apresentará uma descrição do PLN.

3.1. História

O Processamento de Linguagem Natural surgiu praticamente junto com os primeiros sistemas computacionais elaborados, entretanto, acredita-se que um dos maiores propulsores para a área da lingüística computacional tenha sido a guerra fria, onde o uso da criptografia foi bastante utilizado.

Em 1943, sobre a liderança de Allan Turing, um grande matemático e teórico da computação, uma equipe projetou um computador chamado Colossus que foi utilizado no final da Segunda Guerra Mundial. O Colossus tinha a missão de quebrar códigos alemães ultra-secretos produzidos com uma máquina de criptografia chamada Enigma. Em 1949, o vice-presidente da fundação Rockeffeler, que era um grande conhecedor de trabalhos sobre a criptografia computacional, escreveu e distribuiu 200 cópias de uma carta que convidava universidades e empresas a desenvolverem projetos no campo da tradução automática. Essa carta continha algumas teorias, metodologias e problemas que deveriam ser considerados nesses estudos, como por exemplo o problema da polissemia. Os projetistas de PLN da época acreditavam que não havia distinção entre traduzir automaticamente linguagens naturais e decifrar códigos, mas conforme novos estudos surgiam, essas distinções iam se tornando mais claras.

Em 1954 houve uma demonstração, na Universidade de Georgetown, de um sistema capaz de traduzir cinqüenta frases selecionadas do russo para o inglês. O dicionário construído continha 250 palavras e a gramática possuía seis regras. Esse sistema acabou fazendo sucesso e atraiu novos financiadores pelo mundo e desde então, os projetos envolvendo tradução se multiplicaram até chegar, em 1966, ao total de mais de 20 milhões de dólares gastos, mas com poucos resultados relevantes.

No final da década de 60, os grandes gastos já eram questionados com base nos poucos avanços obtidos até então, entretanto, em 1970, um doutorando chamado Winograd desenvolveu em sua tese no MIT um sistema computacional que simulava graficamente, no monitor de um computador, o braço de um robô que manipulava um conjunto de blocos sobre a superfície de uma mesa. Através de instruções, em inglês, digitadas no teclado do computador, o robô executava os comandos inseridos. Esse projeto foi um grande marco e possibilitou uma real demonstração da interação homem-máquina e a PNL, novamente, apareceu como uma grande promessa. Desde então os avanços continuaram aparecendo.

3.2. Descrição

O PLN é um subcampo da Inteligência Artificial (e da Lingüística) que estuda a compreensão e geração automática de linguagem natural. Os sistemas capazes de geração de linguagem natural utilizam informações existentes em bancos de dados de computadores em linguagem compreensível aos humanos. Os sistemas capazes de compreender linguagem natural convertem–na em representações mais formais, melhor manipuláveis pelos computadores.

O PLN é uma tarefa altamente complexa por envolver estudos nas áreas de Ciência da Computação, Lingüística e Ciências Cognitivas.

Os computadores possuem a capacidade de compreender instruções em linguagens de programação, como é o caso do java, delphi, c, pascal; entretanto, não são capazes de processar com naturalidade as linguagens humanas. As linguagens de programação diferem das linguagens naturais por conterem regras fixas e estruturas lógicas invariáveis e, dessa forma, o computador processa matematicamente cada comando. Nas linguagens naturais existem ambigüidades, nuances e diversas interpretações que dependem do contexto, do conhecimento do mundo, cultura, conceitos abstratos, etc. Segue abaixo um exemplo:

(5) Luana viu o menino com a luneta.

A sentença (5) possui uma ambigüidade que gera problemas para o processamento, visto que não podemos chegar a qualquer conclusão se não houver contexto. Não sabemos se é Luana que está com a luneta ou se é o menino. Esse é um problema que só pode ser resolvido analisando-se com o contexto da frase. Abaixo segue outro exemplo:

(6) Gosto de maçã.

Observando o exemplo (6), podemos verificar que “gosto” pode ser um substantivo ou a 1ª pessoa do verbo gostar. Seguindo o exemplo (5), essa sentença só pode ser desambiguizada a partir de uma análise semântico/pragmática.

Processar a linguagem natural é fornecer ao computador a capacidade para “entender” as regras da linguagem natural, assim como as ambigüidades, conceitos abstratos, interpretar sentidos e assim por diante. Os sistemas que existem hoje ainda estão longe de entender conceitos abstratos,

“aprenderem” novos conceitos e ser capazes de se tornar autodidatas. Claro que os sistemas existentes já são bastante avançados se os compararmos com os sistemas criados há dez anos, mas não é possível saber se poderão, um dia, alcançar o nível humano de processamento de linguagem natural.

Outros problemas contextuais podem ser observados na utilização dos pronomes, como demonstra o exemplo abaixo:

(7) O telhado caiu sobre Carlos. Ele machucou os pés.

Nesse exemplo, invariavelmente percebemos que o pronome na segunda frase se refere a “Carlos”, porque telhado não possui pés, mas se substituíssemos “telhado” por “menino”, poderíamos observar a existência da ambigüidade. Os computadores são incapazes de lidar com essas nuances caso não possuam conhecimento semântico e pragmático.

4. O SISTEMA PROPOSTO

Neste capitulo, serão abordados os conhecimentos necessários para o desenvolvimento do sistema proposto nesta dissertação. No item 4.1, será feita uma breve introdução sobre a construção do sistema. O item 4.2 apresenta uma descrição sobre a elaboração do método. A arquitetura proposta pelo sistema será descrita no item 4.3. Em seguida, o item 4.4 apresenta a elaboração dos bancos de dados desenvolvidos nessa dissertação, e detalha a arquitetura do sistema com base na arquitetura cliente/servidor. No item 4.5, é abordada a metodologia adotada para o desenvolvimento do sistema e das bases de dados. Nos itens subsequentes (4.6, 4.7, 4.8, 4.9 e 4.10), serão descritos os módulos gerenciador, módulo de memória recente, módulo de análise léxica, módulo de análise morfológica e o módulo de análise sintática respectivamente.

4.1. Introdução

O sistema proposto nesta dissertação possui uma arquitetura voltada para internet. Dessa forma, pode ser utilizado por qualquer pessoa que possua acesso à rede mundial de computadores. Ao acessar a página principal do sistema, haverá uma tela com uma apresentação sucinta e um campo onde o usuário poderá inserir um texto. Ao inserir um texto e pressionar o botão “processar”, o sistema processará o texto inserido e retornará uma nova página para o usuário com o texto inserido e em destaque os substantivos e verbos contidos no texto.

4.2. A elaboração do método

Através de estudos realizados na área de PLN e na área de lingüística, algumas idéias começaram a surgir para a elaboração um método capaz de dar conta dos complexos mecanismos das linguagens naturais.

Primeiramente, apareceram idéias simples que foram rapidamente descartadas, como a idéia de fazer **apenas** uma busca simplificada em um banco de dados que contivesse todos os vocábulos de uma determinada linguagem natural, em nosso caso, o Português do Brasil (PB). Nessa idéia, cada vocábulo do PB seria cadastrado em um BD juntamente com sua classe gramatical. Dessa forma, por exemplo, o vocábulo "homem" seria cadastrado como substantivo, gênero masculino, número singular e o sistema precisaria apenas fazer uma consulta ao BD para recuperar a classe gramatical.

Esse seria o cenário perfeito, entretanto esse modelo foi descartado, pois há determinados elementos que possuem ambigüidade categorial, como demonstra o exemplo abaixo:

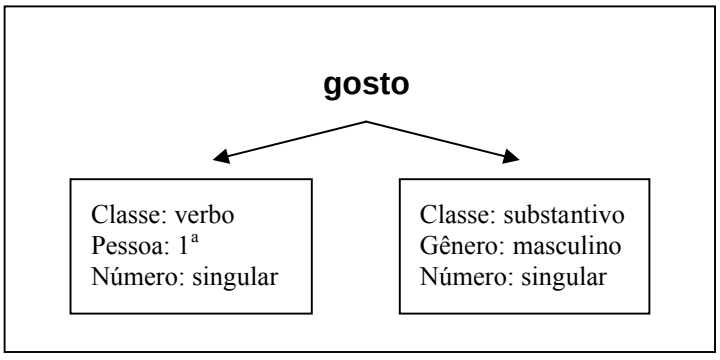

FIGURA 4.1 – EXEMPLO DE AMBIGÜIDADE CATEGORIAL.

Na figura 4.1, podemos notar que um vocábulo com ambigüidade categorial não possui por si só as informações necessárias para sua desambiguação, sendo necessário um contexto para isso. Há ocorrências em que a ambigüidade vai além de duas categorias, como é o caso do vocábulo “a”, que pode ser classificado como artigo, preposição, pronome, dentre outras. Observando a ambigüidade categorial, conclui-se que o sistema possuiria uma complexidade maior do que uma simples busca por uma chave (vocábulo a ser buscado) e um valor (uma classe gramatical). Portanto, esse tipo de busca simplificada em banco de dados não poderia ser o foco desse trabalho, pois o sistema teria que ser capaz de lidar com os casos de ambigüidade categorial semelhantes ao descrito anteriormente.

Durante as buscas por métodos mais efetivos, optou-se pela idéia de desenvolver um sistema que pudesse conter regras de decisão para determinar a classe gramatical dos vocábulos que possuíssem ambigüidade categorial.

Observemos o título da seguinte música interpretada por Cássia Eller:

(8) O gosto do amor.

No exemplo acima, se isolarmos o vocábulo “gosto”, podemos perceber a ambigüidade categorial apresentada pelo mesmo, que pode ser verbo ou substantivo. O sistema proposto seria incapaz de identificar precisamente em que caso se enquadra o vocábulo “gosto” se não fosse por um conhecimento que pudesse ser consultado, ou seja, um conjunto de regras. Com esse embasamento, ao localizar um vocábulo que se enquadre em mais de uma categoria gramatical, o sistema irá processar um conjunto de regras de decisão. As regras acessadas pelo sistema foram criadas em arquivos separados, para que pudessem ser manipulados e atualizados por pessoas que não possuíssem conhecimentos em informática. A idéia foi estabelecer uma fronteira precisa entre

as regras e o sistema em si, ou seja, o código-fonte do sistema não precisaria ser modificado a cada vez que fosse criada uma nova regra de decisão ou que, eventualmente, fosse feita alguma atualização em qualquer uma das regras.

Um dos fatores decisivos para a escolha da linguagem de programação que seria utilizada para o desenvolvimento do sistema foi a capacidade que a linguagem teria para lidar com a incorporação de novos módulos, caso fosse necessário. Em conseqüência disso, optou-se pela linguagem de programação *Java.* Um dos fatores mais significativos para essa escolha foi o fato de que o *Java* é uma linguagem orientada a objetos, o que facilita a modularidade desejada. Além disso, as aplicações desenvolvidas em Java possuem alta portabilidade, sendo possível, com facilidade, migrá-las para sistemas operacionais diferentes, aparelhos portáteis, celulares, dentre outros.

A principal característica das linguagens orientadas a objeto é a decomposição do sistema em objetos com tarefas específicas. A comunicação entre esses objetos é feita através de trocas de mensagens. Essa decomposição facilita a reutilização de classes, assim como a manutenção do sistema.

Para ficar clara a distinção entre classe e objeto, analisemos o exemplo abaixo:

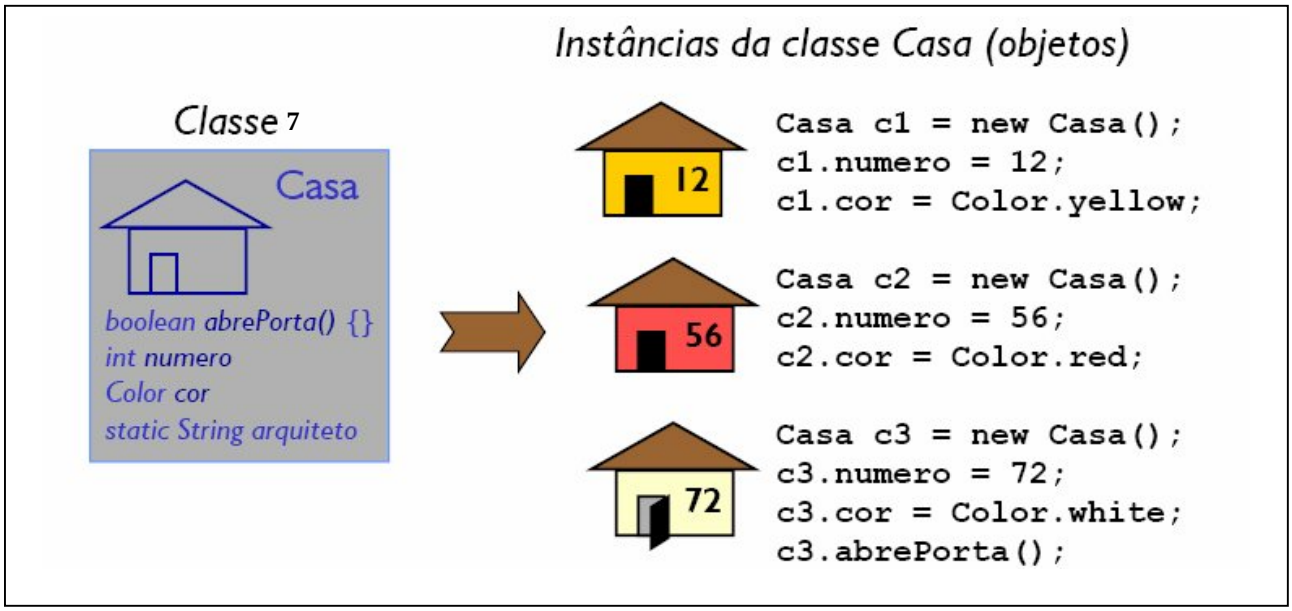

FIGURA 4.2 – EXEMPLO DE CLASSE E OBJETO

Na figura acima, a classe Casa (à esquerda) permite a criação de (n) objetos do tipo casa (à direita). Cada casa criada possui atributos específicos (número, cor, etc.). As classes são os modelos de objetos, são a especificação para um objeto. Em uma analogia, a classe seria a planta da casa, a partir da qual diversas casas poderiam ser construídas. Esse tipo de programação facilita o acoplamento de novas classes ao sistema, além de facilitar sua divisão em módulos que possuem tarefas específicas, como por exemplo, um módulo responsável por analisar morfologicamente uma palavra.

---

[7] AbrePorta() é um método responsável por abrir a porta da casa. Toda casa vai herdar esse método. Toda casa tem um número, representado por um inteiro. Todas as casas terão uma cor específica e terão um arquiteto. Todas os objetos "casa" criados, criadas possuirão esses atributos especificados na classe "casa".

Todo o sistema foi preparado modularmente, para que novos módulos pudessem ser integrados e para que, em caso de futuras melhorias em algum módulo, o mesmo pudesse ser focado, não sendo necessário modificar todo o sistema.

Um exemplo claro disso é o fato de que no inicio do desenvolvimento do SOFIA, o módulo de memória recente não havia sido previsto. Entretanto, a modularidade utilizada permitiu a integração do novo módulo com facilidade. O módulo de memória recente possui uma importância fundamental para o sistema e será descrito mais adiante.

4.3. Arquitetura e funcionamento do sistema proposto

O SOFIA utiliza a arquitetura cliente/servidor. Na arquitetura em questão, o sistema não precisa estar presente no computador do usuário, ou seja, os usuários do sistema não precisam instalar qualquer aplicativo em sua máquina, pois todo o acesso é feito através da internet; todo o código do sistema se encontra no servidor. Essa arquitetura foi escolhida para proporcionar um fácil acesso de qualquer parte do mundo e prover uma maior segurança aos usuários que, muitas vezes, não instalam um determinado programa por precaução contra os vírus de computador.

Dentro da área da informática, um servidor *web* é um sistema computacional designado para fornecer serviços a uma rede de computadores, como por exemplo, páginas *web*, serviço de correio eletrônico, transferência de arquivo, acesso remoto, dentre outros.

Os computadores que solicitam os serviços providos pelo servidor são denominados clientes. O mesmo termo “cliente” é utilizado para designar o usuário que acessa os serviços do servidor.

A arquitetura cliente/servidor é um modelo computacional que separa os clientes dos servidores. Nessa arquitetura, geralmente, os clientes são interligados ao servidor através de uma rede de computadores, como demonstra a figura abaixo.

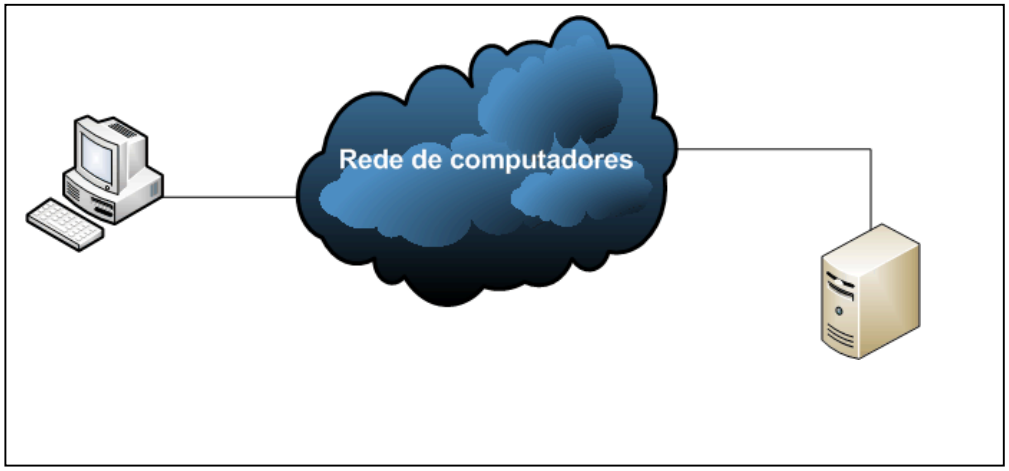

FIGURA 4.3 – ARQUITETURA CLIENTE/SERVIDOR

Os navegadores *web* (ex: internet explorer) são programas que habilitam os usuários (clientes) a interagirem com os *websites* hospedados em um servidor *web*. O navegador *web* também é conhecido como *web browser* ou apenas *browser*, que são termos provenientes da língua inglesa. Em inglês, *browser* provém do verbo *to browse*, que significa olhar páginas em um livro, revista, etc.

Existe a intenção de, futuramente, criar um módulo onde os usuários possam interagir com o sistema, mesmo que seja apenas para definir a classificação realizada como correta ou incorreta. Isso será feito para que o sistema possa evoluir significativamente com o tempo. Caso o sistema

estivesse instalado nos computadores dos usuários, o envio dessas informações seria mais problemático.

A figura 4.3 apresenta uma descrição detalhada da arquitetura proposta para o SOFIA.

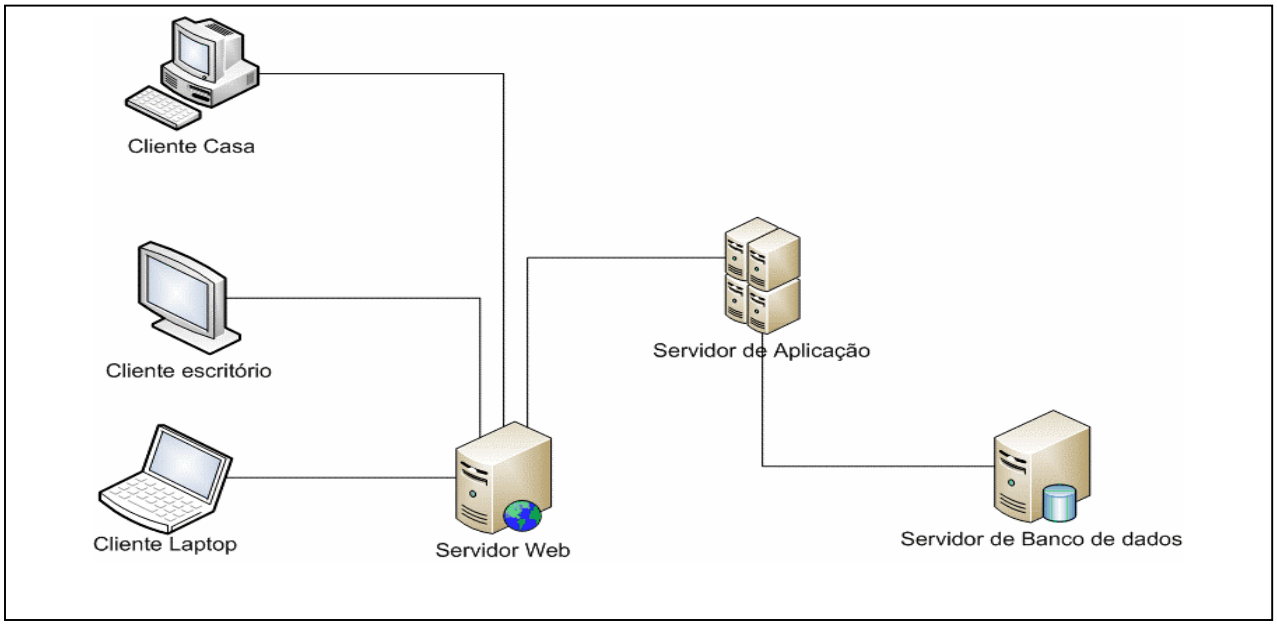

FIGURA 4.4 – ARQUITETURA MACRO DO SISTEMA

O usuário utiliza um navegador *web* e através do mesmo faz o acesso ao servidor *web*. O servidor *web* é responsável por servir às solicitações feitas pelos usuários; é o responsável por fazer a interligação entre o usuário e o servidor de aplicação. Quando o usuário insere um texto no *navegador web* e envia uma solicitação de processamento, o *servidor web* envia essa solicitação para o SOFIA (presente no servidor de aplicação) que poderá acessar o BD e retornar o resultado obtido para o *servidor web*. Por fim, o *servidor web* retorna para o usuário a página com o resultado obtido.

Abaixo é apresentado um esquema demonstrando a arquitetura interna do SOFIA. Pode-se ver dois módulos principais, um que contêm a frase a ser inserida e é descrito como “navegador *web*” e outro que é descrito como “servidor *web*” e é composto, basicamente, pelo módulo gerenciador, módulo de memória recente, módulos de análise léxica, morfológica e sintática.

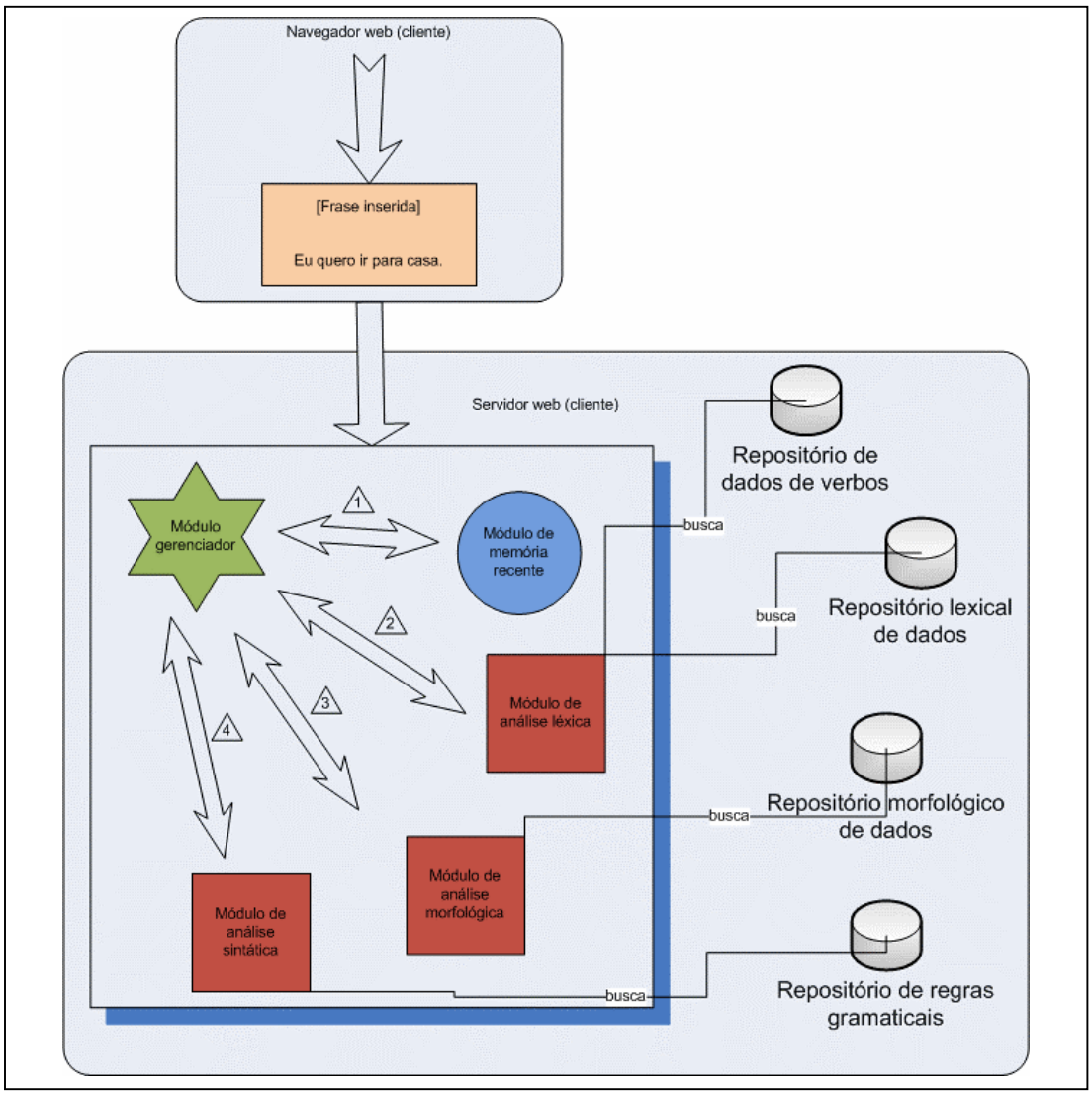

FIGURA 4.5 – ARQUITETURA INTERNA DO SISTEMA

Atualmente, o sistema encontra-se disponível na internet através do endereço http://www.projetosofia.com.br .

A figura abaixo apresenta um resultado obtido através da interface do sistema:

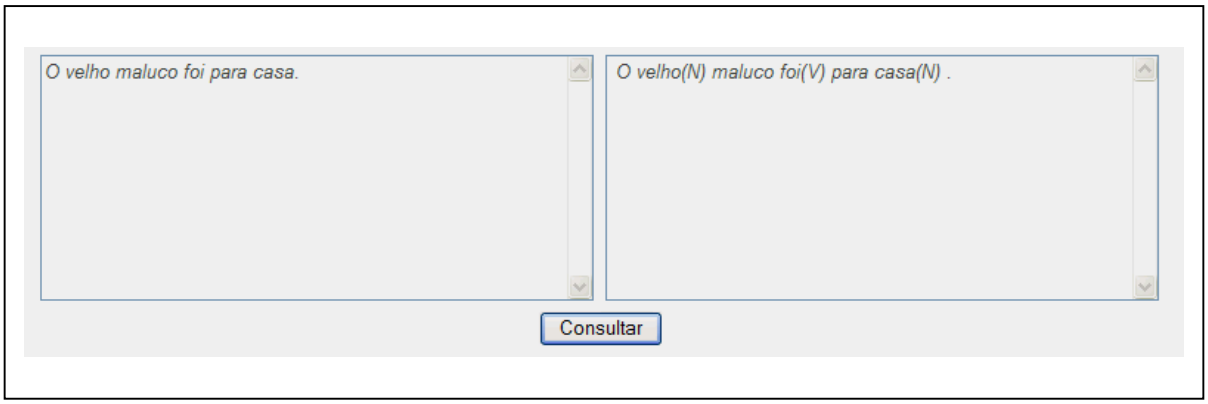

FIGURA 4.6 – RESULTADO OBTIDO PELO SISTEMA

A figura apresenta o sintagma “o velho maluco” em que velho foi classificado como nome em função da característica mais prototípica de ordem vocabular ter sido adotada pelo sistema, ou seja, a precedência do nome ao adjetivo.

O sistema irá processar todos os itens com ambigüidade categorial e irá categorizar cada vocábulo com apenas uma classe gramatical, representando os nomes com (N) e os verbos com (V), como demonstra a figura 4.6.

4.4. Elaboração de base de dados lexicais

O sistema proposto necessitava de uma base de dados que servisse de apoio para a identificação das possíveis classes gramaticais de um vocábulo e, portanto, estabeleceu-se a necessidade de construir um banco de dados capaz de prover essas informações.

A elaboração do banco de dados envolveu um grande esforço de cadastro e processamento de palavras do Português do Brasil. Atualmente, mais de 500.000 entradas e suas respectivas classes gramaticais foram inseridas no banco de dados. Durante a geração do banco de dados, todas as classes gramaticais de cada vocábulo foram cadastradas.

Os bancos de dados elaborados possuem apenas informações de ordem morfossintática, já que neste projeto ainda não há possibilidade de processamento semântico das sentenças, conforme mencionado. Durante a elaboração do projeto, dedicou-se uma grande parte do tempo para a criação de bancos de dados massivos que contivessem um baixo índice de erros. Por esse motivo, a criação dos bancos de dados foi baseada nos dicionários Michaelis (http://michaelis.uol.com.br/moderno/portugues/) e Houaiss eletrônico v. 1.0 (2001).

Para essa primeira versão do sistema, não houve cadastramento de locuções como “uma vez que” ou “o que quer que seja”, sendo apenas cadastradas as formas de palavras de radical único (cabeça) e hifenizadas (pé-de-cabra); entretanto, nada impede que tais locuções ou formas compostas sejam futuramente cadastradas e tratadas pelo sistema.

O sistema contempla os vocábulos homônimos, mas apenas uma entrada é apresentada no BD, ou seja, o sistema apresenta **uma** entrada com (1 - n) classes gramaticais. Segue abaixo a imagem de uma consulta feita pelo sistema ao banco de dados, inserindo como entrada o vocábulo “velho” e o vocábulo “gosto”.

```
vocábulo:  velho
Classe(s): [ADJ_M_S, N_M_S]

vocábulo: gosto
Classe(s): [N_M_S]
           [[gostar] V_PRES_IND_1_S]
```

FIGURA 4.7 – EXEMPLO DE BUSCA EFETUADA PELO SISTEMA

A figura acima demonstra a relação de (1 → n) entre vocábulo e classes. Um vocábulo pode possuir (n) classes. "Velho" pode ser um adjetivo do gênero masculino e número singular (ADJ_M_S), assim como um substantivo masculino no singular (N_M_S). Já o vocábulo "gosto" pode ser um substantivo masculino no singular (N_M_S) ou um verbo no presente do indicativo da 1ª pessoa do singular (V_PRES_IND_1_S).

Os vocábulos modificados em gênero, número e grau não foram cadastrados nos BD, ou seja, ao analisarmos as possíveis ocorrências para o adjetivo "engraçado", podemos ver que não constam no BD os vocábulos "engraçada", "engraçadas" ou "engraçadinha". O sistema utilizará o módulo de análise morfológica para obter as classes gramaticais para os vocábulos flexionados. Os vocábulos flexionados só se encontrarão no BD quando os mesmos possuírem significados semânticos distintos do vocábulo não flexionado, como é o caso de "simpáticos", que possui a classe gramatical de adjetivo modificado em número e também pode ser substantivo masculino. O esforço para o cadastro de todos os adjetivos e substantivos modificados seria excessivo e, por esse motivo, não foram realizados no BD. Caso o cadastramento desses vocábulos fosse feito de forma automática a partir de regras, poderia gerar registros incorretos ao se deparar com as tantas exceções existentes no PB.

O sistema acessa alguns repositórios de dados que foram criados seguindo alguns critérios de arquitetura. Primeiramente, estabeleceu-se que quaisquer informações de regras deveriam ser cadastradas em arquivos que pudessem ser manipulados por pessoas que não possuíssem conhecimentos em programação de sistemas. Dessa forma, optou-se por cadastrar siglas, prefixos, sufixos, regras de gênero, número e grau, regras sintáticas e morfológicas em arquivos XML, que serão descritos mais adiante.

Abaixo podemos visualizar de uma forma mais transparente a separação entre os repositórios e o número de registros inseridos em cada um deles:

- Repositório de verbos (Banco de Dados contendo 5924 verbos) conjugados, o que se resume a mais de 330.000 (337493) diferentes entradas.
- Repositório lexical de dados (Banco de Dados contendo 253.792 entradas **não-verbos**)
- Repositório de siglas contendo 14 siglas
- Repositório de abreviações contendo 14 abreviações
- Repositório de sufixos contendo 2 sufixos[8] (apenas os formadores de advérbio)
- Repositório de regras de número contendo 31 regras
- Repositório de regras de aumentativos contendo 23 regras
- Repositório de regras de diminutivos contendo 6 regras
- Repositório de regras de gênero contendo 9 regras
- Repositório de prefixos contendo 42 prefixos com hífen como (“pré-“)

[8] As terminações cadastradas foram “–amente” (calmo/calmamente) e “–mente” (amável/amavelmente).

- Repositório de prefixos contendo 3 prefixos sem hífen como (“des”)
- Repositório de preposições contendo 127 preposições
- Repositório de regras sintáticas contendo 64 regras

A lista com os 5924 verbos foi retirada do corretor gramatical CoGrOO[9]. A mesma lista foi utilizada para, através de consultas aos dicionários Michaelis e Houaiss, cadastrar toda conjugação de cada um dos verbos no BD verbal. Diferente dos substantivos, os verbos foram cadastrados flexionados para que o sistema pudesse obter uma confiabilidade maior ao consultar esse BD. Ainda que os verbos regulares sigam alguns paradigmas, nenhum módulo de processamento e/ou tratamento desses paradigmas foi desenvolvido para o sistema proposto. Ao cadastrar todos os verbos flexionados, evita-se que vocábulos como “mares”, plural do substantivo “mar” possam ser confundidos com verbos flexionados que possuem a terminação “ares”, como seria o caso de “amares”, verbo da segunda pessoa do singular do presente do subjuntivo.

O repositório lexical e o repositório verbal foram cadastrados utilizando-se o Banco de Dados Mysql[10]. Entretanto, durante o desenvolvimento do sistema, observou-se que as consultas ao Mysql começaram a se tornar lentas, dado o grande número de registros e consultas que estavam sendo efetuadas. Pesquisou-se outras tecnologias e, visto que as (1 - n) buscas efetuadas pelo sistema eram bastante custosas, optou-se pelo banco de dados JDBM[11], tecnologia desenvolvida para

---

[9] Corretor gramatical do Open Office.

[10] http://www.mysql.com/

[11] http://jdbm.sourceforge.net/

linguagem *Java* e que se comportou de forma mais performática. Ambos os BD (lexical não-verbal e verbal) foram migrados para o JDBM.

Todos os demais repositórios foram cadastrados em arquivos XML[12]. O XML é um formato que dá suporte a criação de documentos com dados organizados de forma hierárquica e possui portabilidade, visto que diversas linguagens de programação são capazes de lidar com esse tipo de arquivo. Segue abaixo as abreviações cadastradas no sistema, presentes no formato XML, para "U$", "R$", e "km".

```
1  <?xml version="1.0" ?>
2  <siglas>
3     <siglas-list>
4        <sigla id="1">
5           <nome>US$</nome>
6           <classe>SIG_MONET</classe>
7        </sigla>
8        <sigla id="2">
9           <nome>R$</nome>
10          <classe>SIG_MONET</classe>
11       </sigla>
12       <sigla id="3">
13          <nome>km</nome>
14          <classe>SIG_ESPAC</classe>
15       </sigla>
16    </siglas-list>
17 </siglas>
```

FIGURA 4.8 – ARQUIVO XML DE ABREVIAÇÕES

No arquivo XML apresentado acima, utilizou-se o rótulo "SIG_MONET" para identificar uma sigla monetária e "SIG_ESPAC" para identificar uma sigla espacial.

---

[12] http://www.w3.org/XML/

É importante notar que qualquer pessoa que não possua conhecimentos em programação poderá cadastrar novas siglas. A idéia é exatamente que esses repositórios possam ser atualizados constantemente.

Um dos repositórios com mais importância no sistema é o repositório de abreviações, que são utilizadas para que o sistema seja capaz de fazer distinção entre um ponto final e um ponto contido em uma abreviação, como é o caso da abreviação utilizada para "senhor" (sr.). A elaboração desse repositório surgiu com o aparecimento de erros na segmentação de frases com abreviações, visto que os pontos eram confundidos com pontos finais de frase. O repositório de dados relativo às abreviações foi gerado de acordo com buscas na internet.

4.5. Metodologia

A metodologia utilizada para o desenvolvimento de um analisador morfossintático capaz de reconhecer nomes e verbos em sentenças do Português do Brasil utiliza informações provenientes de repositórios de dados, os mesmos que foram descritos no item 4.4.

Além disso, algumas classes de palavras do PB serão incluídas em um módulo do sistema denominado módulo de memória recente. As classes incluídas nesse módulo são as classes fechadas de palavras, ou seja, os pronomes, artigos, preposições, etc. Essas informações serão inseridas em memória para que seu acesso ocorra mais rapidamente. Uma consulta ao banco de dados é mais custosa para o sistema do que uma consulta à memória. Como esses vocábulos inseridos no módulo de memória recente fazem parte de um grupo finito e bastante utilizado de palavras, o módulo foi construído para que o sistema apresentasse uma melhor performance.

Além dos repositórios utilizados para a identificação das classes gramaticais dos termos de uma oração, o sistema proposto necessitaria de uma base de dados para efetuar os testes. A base de dados utilizada para os testes realizados com o sistema foi o banco de dados do CETEN-Folha, que possui aproximadamente 24 milhões de palavras e foi obtida a partir de 365 edições do jornal brasileiro Folha de São Paulo no ano de 1994.

Antes de iniciar a construção dos códigos do SOFIA, fez-se uma análise sobre a melhor forma de se modularizar o sistema e optou-se por criar 3 módulos fundamentais: módulos de análise lexical, morfológica e sintática.

Ainda que o sistema possua módulos para cada uma dessas fases, existe uma relação entre eles, ou seja, o sistema acessará cada módulo conforme seja necessário, podendo variar e intercalar os acessos.

As entradas nos bancos de dados são compostas por um vocábulo e suas classes de palavras correspondentes conforme descrito na tabela 4.1.

Classe de palavras
abreviatura
adjetivo
advérbio
artigo
conjunção
preposição
numeral
nome próprio
verbo
interjeição
nome (substantivo)
número
pronome
sigla

TABELA 4.1 – CLASSES DE PALAVRAS UTILIZADAS

Foram utilizados rótulos para cada classe de palavra acima. Dessa forma, a manipulação das classes dentro do sistema poderia se tornar mais simplificada. A tabela 4.2 apresenta as listas de abreviações utilizadas.

Classe	**Sigla**
abreviatura	[ABREV]
adjetivo	[ADJ]
advérbio	[ADV]
artigo	[ART]
conjunção	[CONJ]
preposição	[PREP]
numeral	[NUM]
nome próprio	[NP]
verbo	[V]
interjeição	[INTERJ]
nome	[N]
número	[NUMERO]
pronome	[PRON]
sigla	[SIG]

TABELA 4.2 – RÓTULOS DE ENTRADAS NO BD

Outros rótulos utilizados pelo sistema poderão ser encontrados no Apêndice A, ao final da dissertação. A tabela 4.3 apresenta um exemplo de como as entradas se apresentam no banco de dados. Embora apareçam etiquetas de nomes e verbos na mesma célula da tabela, é importante ressalvar que existe um banco de dados separado para os verbos, ou seja, a tabela abaixo se apresenta em um nível ilustrativo.

Palavra	Classe
casa	[N_F_S] [[casar] V_PRES_IND_3_S]
carro	[N_M_S]
começo	[N_M_S] [[começar] V_PRES_IND_1_S]
bonito	[ADJ_M_S] [ADV] [INTERJ] [N_M_S]

TABELA 4.3 – EXEMPLO DE ENTRADA NO BD

Dentro do léxico, há grupos de regras que caracterizam o comportamento de um subconjunto de vocábulos da linguagem, como o caso de diminutivos de substantivos masculinos terminados em "inho". O sistema possui um conjunto de regras para reconhecer tais comportamentos, freqüentes no PB. Essas regras fazem parte do módulo relativo à análise morfológica, que será descrito mais adiante.

Todas as entradas no BD lexical possuem informações gramaticais. Quando um vocábulo estiver enquadrado em mais de uma classificação gramatical, como é o caso de "gosto" e "casa", as diversas possibilidades serão associadas ao mesmo. Abaixo será descrito o tratamento efetuado para cada classe gramatical presente no sistema.

4.5.1 Artigo

A classe dos artigos é constituída por um número limitado de palavras e, esse é um dos motivos pelos quais foram inseridos no módulo de memória recente. Este módulo será detalhado mais adiante. O sistema inclui informações morfológicas de gênero, número e tipo (definido ou indefinido) para os artigos.

A ocorrência de artigos, sejam os definidos ou indefinidos, são de grande importância para o sistema, visto que, em muitos casos, delimitam uma série de elementos que podem precedê-los (ex: verbos) ou sucedê-los (ex: nomes). Por se encontrarem no módulo de memória recente, não é necessário fazer consultas ao banco de dados a cada vez que os mesmos são encontrados. Isso é de suma importância para o sistema, pois, conforme mencionado anteriormente, as consultas ao banco de dados demandam mais tempo do que as consultas à memória.

4.5.2 Substantivo

No Português, os substantivos apresentam informações morfológicas de gênero, número e grau. Conforme mencionado no item 4.4, as informações de gênero (ex: menina), número (ex: meninos) e grau (ex: menininho) dos substantivos não foram inseridas no BD lexical, pois as entradas são apresentadas como em um dicionário. As informações de gênero, número e grau serão processadas pelo módulo de análise morfológica.

Casos como “caminho” não serão identificados como diminutivo, visto que o módulo de análise morfológica só é empregado quando os vocábulos não são encontrados no banco de dados lexical e o vocábulo “caminho” será encontrado no BD lexical e no BD de verbos.

A formação do grau dos substantivos pode ser dividida em aumentativo sintético, aumentativo analítico, diminutivo sintético e diminutivo analítico. Serão tratados nesse trabalho apenas o aumentativo sintético e o diminutivo sintético. Quando algum substantivo no gênero masculino ou no número plural for inserido como entrada, será, primeiramente, buscado no BD sem qualquer processamento morfológico, pois existem alguns substantivos femininos e substantivos plurais cadastrados. Caso não sejam encontrados, o módulo de processamento

morfológico será acionado para tentar recuperar informações morfológicas e efetuar as transformações necessárias para descobrir o vocábulo conforme foi cadastrado no BD. Esse processo será explicado com mais detalhes no item 4.9.

4.5.3 Adjetivo

O adjetivo, assim como o substantivo, pertence a uma classe aberta de palavras. Possuem informações morfológicas de gênero, número e grau. No BD lexical não constam informações relativas ao gênero, número e grau dos adjetivos, assim como ocorre para os substantivos. Essas informações serão processadas pelo módulo de análise morfológica, que contêm regras para a identificação de possíveis ocorrências de vocábulos flexionados.

O processamento realizado pelo módulo de análise morfológica será descrito com detalhes no item 4.9. Esse módulo funciona de forma idêntica para os substantivos e os adjetivos. Os possíveis problemas de ambigüidade categorial entre adjetivos e substantivos também estão descritos no item 4.9.

4.5.4 Pronome

Os pronomes possuem informações morfológicas de gênero, número e pessoa gramatical. Os pronomes pessoais oblíquos, os pronomes pessoais retos, pronomes relativos, possessivos, demonstrativos e pronomes indefinidos foram cadastrados no módulo de memória recente para um acesso mais rápido. Assim como os pronomes pessoais, os pronomes possessivos possuem características que os identificam como pessoa conforme segue abaixo:

Pessoa	Pessoal	Possessivo
1ª pessoa	*eu, nós*	*meu(s), nosso(s)*
2ª pessoa	*tu, você(s), vós*	*teu(s), vosso(s)*
3ª pessoa	*ele, eles*	*dele, deles, seu(s)*

TABELA 4.4 – PRONOMES POSSESSIVOS

Os pronomes fornecem um grande auxílio ao sistema, visto que muitas vezes a desambiguização de constituintes pode ser caracterizada pela presença desses elementos, como é o caso do exemplo abaixo:

(9) "Eu gosto de chocolate."

O constituinte "gosto" será classificado como verbo, já que a ocorrência de um pronome pessoal precedendo imediatamente um verbo é o padrão mais recorrente.

4.5.5 Outras classes gramaticais

As preposições, advérbios, conjunções, interjeições são classes gramaticais invariáveis e encontram-se cadastradas no BD. Alguns sufixos foram cadastrados em repositórios e são acessados pelo módulo de processamento morfológico, assim como o sufixo (-mente). Em Português, é possível formar corretamente um advérbio em "-mente" a partir de adjetivos (ex: calmo → calmamente), desde que algumas regras sejam respeitadas. O sistema é capaz de reconhecer advérbios terminados em "–mente", mesmo que os mesmos não estejam cadastrados no BD. Caso encontre um advérbio terminado em "-mente", primeiramente, o sistema efetua uma

busca no BD para verificar a existência da palavra. Caso a mesma não seja encontrada, o módulo de análise morfológica é acessado. A partir desse momento, diversas regras serão processadas, e uma dessas regras é responsável por verificar a ocorrência do sufixo "-mente". É importante notar que existem casos de verbos terminados em "mente", como é o caso de "aumente". Esses verbos ou possíveis nomes terminados em "mente" não constituem um problema para o sistema, visto que seriam encontrados no BD e o módulo de análise morfológica não chegaria a ser acessado. Segue abaixo o resultado obtido pelo sistema ao consultarmos o vocábulo "aumente".

```
vocábulo: aumente
Classe(s): [aumentar] V_PRES_SUB_3_S
           [aumentar] V_PRES_SUB_1_S
```

FIGURA 4.9 – CONSULTA DO VOCÁBULO "AUMENTE"

Podemos perceber pela figura acima que embora o vocábulo "aumente" possua a terminação "mente", o sistema retornou apenas a classe gramatical verbo. O sistema só faz o processamento morfológico que processa o sufixo "mente" se não encontrar um verbo com essa terminação. O funcionamento interno do módulo de análise morfológica será descrito com mais detalhes no item 4.9.

4.6. Módulo gerenciador

Este módulo é o controlador central do SOFIA, ele é o responsável por enviar todas as ordens ao sistema, processar os retornos obtidos e, por conseguinte, enviar ou não uma nova ordem. Assim que um texto é inserido como entrada para o sistema e submetido, esse módulo é acionado para receber o texto e gerenciar todo o processamento, desde a segmentação das frases até o resultado que é exibido para o usuário.

Sempre que uma frase for processada, o módulo de memória recente será acessado, assim como o módulo de análise léxica. Entretanto, os módulos de análise morfológica e análise sintática só serão acessados quando houver necessidade. Caso o texto inserido possua apenas uma classe gramatical para cada constituinte, não haverá necessidade de acessar o módulo de análise morfológica, visto que esse módulo só é acessado quando o módulo de análise lexical não consegue recuperar a classe gramatical de um vocábulo no BD lexical. O módulo de analise sintática só será acessado caso um vocábulo possua mais de uma classe gramatical, e em nosso caso acima, o texto inserido possui uma classe gramatical para cada vocábulo, ou seja, não existem ambigüidades categoriais.

Por outro lado, caso exista vocábulos não encontrados no BD, como por exemplo, vocábulos modificados em grau, o sistema acessará o módulo de análise morfológica, assim como, caso exista em qualquer frase um vocábulo com ambigüidade categorial, o módulo de análise sintática será acionado.

A opção por um módulo que gerenciasse o sistema surgiu para que as manutenções e melhorias ocorressem de forma mais eficiente, visto que é possível saber qual parte do sistema é responsável por enviar as ordens.

4.7. Módulo de memória recente

O módulo de memória recente é responsável por carregar alguns vocábulos e suas classes em memória assim que o sistema é iniciado. Dessa forma, sempre que esses termos forem encontrados nos textos inseridos, o sistema não precisará acessar o banco de dados, que é a tarefa que leva mais tempo em uma consulta. O nome "memória recente" foi atribuído ao módulo pelo fato de que esses termos aparecem mais frequentemente e, portanto, o acesso a eles seria mais imediato do que o acesso a um vocábulo que não estaríamos acostumados a utilizar, como por exemplo, os vocábulos virgulta[13] ou ceteno[14]. Foram inseridos os artigos, preposições, pronomes, siglas, abreviaturas, prefixos e sufixos. É importante ressaltar que em alguns casos, como é o caso da preposição "entre", uma consulta também é feita ao banco de dados para que ocorra a busca das ambigüidades gramaticais. Isso também ocorre na inicialização do sistema. A figura abaixo demonstra um trecho do arquivo XML das preposições.

```
<prep id="13" active="true">
    <preposicao checkindb="true">exceto</preposicao>
</prep>
<prep id="14" active="true">
    <preposicao checkindb="true">entre</preposicao>
</prep>
<prep id="15" active="true">
    <preposicao checkindb="true">mediante</preposicao>
</prep>
<prep id="16" active="true">
    <preposicao checkindb="true">para</preposicao>
</prep>
```

FIGURA 4.10 – ARQUIVO XML DE PREPOSIÇÕES

[13] varinha flexível

[14] substância (C2H2O) que constitui gás bastante reativo, esp. us. como agente de acetilação em síntese orgânica (p.ex., na aspirina)

O sistema carrega todas as preposições contidas nesse arquivo para a memória do sistema e, quando encontra o atributo (checkindb=true), faz uma consulta ao BD para recuperar as possíveis classes gramaticais existentes.

Esse é um módulo que pode ser expandido para que possa incluir outros vocábulos de uso freqüente, caso haja uma baixa performance do sistema, como por exemplo, palavras que fossem utilizadas com mais freqüência na linguagem.

4.8. Módulo de análise léxica

Este módulo do sistema é responsável pela segmentação de um determinado texto que é inserido como entrada no sistema, em elementos significativos, denominados *tokens.* Inicialmente, esse módulo segmenta o texto em frases. As frases são delimitadas quando são encontrados os sinais de interrogação (?), exclamação (!), ponto e vírgula (;) e ponto final (.). Conforme mencionado, existe um tratamento para verificação de abreviações (também presentes em arquivo XML), em que o sinal de ponto não se comporta como um ponto final, mas sim como um ponto ao final de uma abreviação.

Posteriormente, essas frases são segmentadas nos *tokens*, como as palavras e os itens de pontuação. Após essa fase, com base em consultas ao léxico, serão atribuídas a cada *token* as possíveis informações de classes de palavras existentes. Caso não haja qualquer informação relevante no banco lexical, o módulo gerenciador aciona o módulo de análise morfológica que irá processar um conjunto de regras para tentar extrair algum traço morfológico relevante para que esse *token* possa, finalmente, ser reconhecido no banco lexical. Como o banco de dados lexical não

possui vocábulos cadastrados no diminutivo e no plural, os *tokens* precisam, portanto, sofrer modificações antes de serem associados ao seu verbete. Dessa forma, o analisador morfológico se apresenta como um módulo auxiliador e será descrito no próximo item.

O método de análise léxica também é conhecido como *scanner*. Essa é uma técnica bastante conhecida em sistemas de computação que se propõem a transformar os códigos escritos em uma linguagem de programação para outra.

4.9. Módulo de análise morfológica

Caso o analisador léxico não encontre qualquer informação relevante no BD lexical, é retornada uma informação para o módulo gerenciador. Em seguida, o módulo gerenciador irá acionar o analisador morfológico como ilustra a figura 4.11.

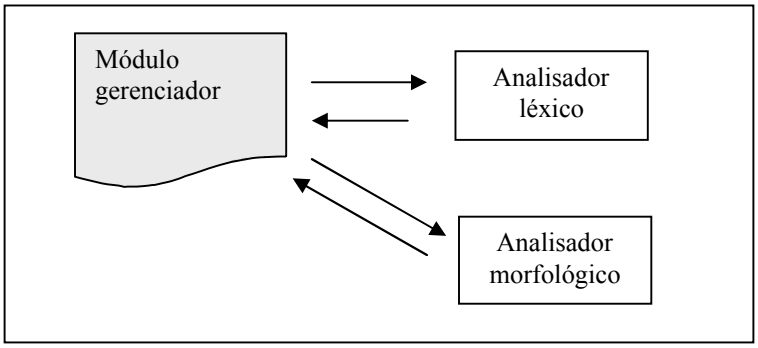

FIGURA 4.11 – ARQUITETURA DO SISTEMA

O módulo de análise morfológica irá realizar regras transformacionais em cima do léxico recebido do módulo gerenciador para tentar reconhecer a morfologia da palavra, baseando-se nas regras morfológicas. Caso reconheça alguma regra morfológica, o módulo irá, então, processar essas informações e proverá para o módulo gerenciador um conjunto de palavras passíveis de serem encontradas no banco de dados lexical. Os processos realizados para o gênero, número e grau dos substantivos e adjetivos serão descritos nos itens 4.9.1, 4.9.2, e 4.9.3. É importante deixar claro que esses processamentos só serão realizados se o vocábulo não existir no BD lexical.

4.9.1 Número dos substantivos e adjetivos

Algumas regras foram implementadas no sistema com a finalidade de reconhecer a ocorrência flexionada de determinado vocábulo no plural. Primeiro, o sistema verifica o sufixo da palavra em questão e compara com as terminações presentes nas regras associadas ao plural. Caso o sistema encontre uma terminação igual, irá realizar uma transformação no vocábulo em questão e enviará para o módulo gerenciador as possíveis ocorrências do vocábulo transformado.

Abaixo segue uma imagem de um trecho do arquivo XML que contém 31 regras de formação de plurais:

```xml
<?xml version="1.0" encoding="ISO-8859-1" ?>
<rules>
    <numero>
        <rule id="1" active="true">
            <plural>ães</plural>
            <singular>ão</singular>
            <exemplo>botão-botões</exemplo>
        </rule>
        <rule id="2" active="true">
            <plural>ais</plural>
            <singular>al</singular>
            <exemplo>policial-policiais</exemplo>
        </rule>
        <rule id="3" active="true">
            <plural>as</plural>
            <singular>a</singular>
            <exemplo>conta-contas</exemplo>
        </rule>
        <rule id="4" active="true">
            <plural>es</plural>
            <singular>e</singular>
            <exemplo>corte-cortes</exemplo>
        </rule>
```

FIGURA 4.12 – ARQUIVO XML DE PLURAIS

É importante lembrar que todas essas informações de singular e plural são inseridas em memória na inicialização do sistema. Sempre que necessário, o módulo de análise morfológica poderá acessar essa lista de plurais.

Se o sistema fizer uma busca no BD lexical pelo vocábulo “projéteis”, nada será encontrado. Isso ocorre porque há pouca informação sobre plurais no BD. Para um léxico mais enxuto, foi adotada a posição de não cadastrar todas as palavras no plural da língua. A metodologia adotada foi a criação de um módulo que fosse capaz de lidar com as regras derivacionais de plural. O sistema então aplicará um conjunto de regras, como ilustrado na figura abaixo e irá tentar transformar o vocábulo para o singular.

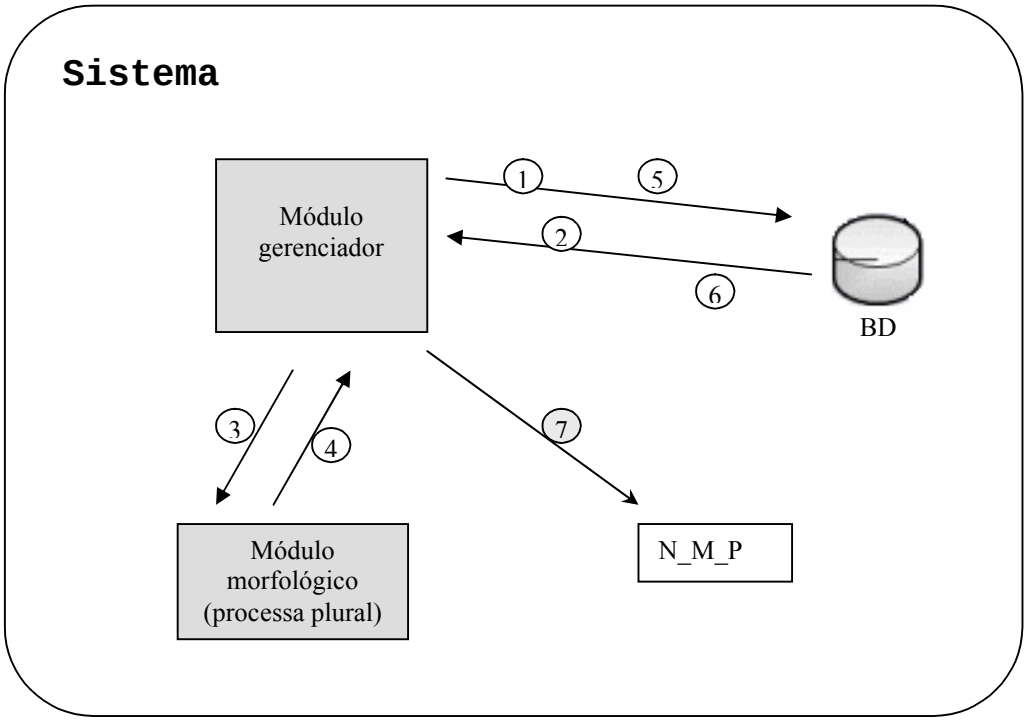

FIGURA 4.13 – ARQUITETURA DO SISTEMA

Para detalhar claramente esse funcionamento, a descrição de cada fluxo, enumerado na figura acima, é apresentada abaixo:

1 – Ao receber o vocábulo “projéteis”, o **módulo gerenciador** fará uma busca no BD.

2- O BD não retorna uma etiqueta para o **módulo gerenciador**, visto que “projéteis” não se encontra cadastrado no BD.

3- O **módulo gerenciador** envia o vocábulo “projéteis” para o **módulo morfológico**.

4- O **módulo morfológico** acessa as regras morfológicas e processa o vocábulo recebido, identificando a terminação de plural. Efetua as transformações necessárias para a recuperação do vocábulo não flexionado e retorna “projétil” para **módulo gerenciador**.

5- O **módulo gerenciador** fará nova busca no BD, agora com a palavra “projétil”.

6- O BD retorna a etiqueta **N_M_S** para o **módulo gerenciador**. O **módulo gerenciador** reprocessa a etiqueta recebida e modifica-a para **N_M_P**.

7- O **módulo gerenciador** provê a etiqueta para o sistema.

Há palavras que poderão derivar para mais de um vocábulo durante as regras transformacionais. Nesse caso, as (n) palavras serão retornadas para o módulo gerenciador e serão buscadas no banco de dados lexical. É importante lembrar que essas regras só serão aplicadas se o vocábulo **não** for encontrado no BD lexical.

A tabela explicita as informações do arquivo XML que contêm as regras de plurais. Para exemplificar o processamento de plurais dentro do módulo morfológico, observemos a 1ª coluna da tabela abaixo. Essa coluna representa as formas nominais de plural.

Plural (entrada)	Singular (saída)	Sufixo plural	Terminação singular
caracóis	caracol	-óis	-ol
amores	amor	-es	-
pudim	pudins	-ins	-im

TABELA 4.5 – EXEMPLO DE ALGUMAS REGRAS DE PLURAL

Caso o usuário entre com uma palavra no plural, por exemplo, “caracóis”, o sistema irá acessar o BD para buscar a palavra em questão. Como a palavra não será encontrada, o módulo morfológico irá consular a lista de terminações no plural, representada pela coluna 3. O módulo morfológico identificará a terminação “-óis” e realizará a transformação na palavra inserida substituindo o sufixo “óis” pelo sufixo correspondente na coluna 4 da tabela, ou seja, a coluna de terminações de singular. Substituindo o sufixo “óis” pelo sufixo “ol”, obtemos a palavra caracol.

Dessa forma, quando o sistema buscar esse “novo” vocábulo no BD (caracol), o mesmo será encontrado. Em todas as palavras com terminações de plural que não forem encontradas no BD sem a utilização do módulo morfológico, o sufixo encontrado na coluna 3 será substituído pelo sufixo da coluna 4.

4.9.2 Gênero dos substantivos e adjetivos

O processamento do gênero ocorre de forma semelhante ao número. As informações de gênero feminino encontram-se em um arquivo XML. Como os registros existentes no BD seguem o padrão de um dicionário, a maioria dos vocábulos é cadastrada no gênero masculino. Analisemos o caso de o sistema receber como entrada o vocábulo “vencedora”. O módulo gerenciador acessará o BD para tentar recuperar as classes gramaticais correspondentes ao vocábulo citado. Visto que o retorno do BD será vazio, o módulo gerenciador acessará o módulo morfológico. O módulo morfológico processará o vocábulo “vencedora”.

O processamento do vocábulo irá derivar para duas saídas distintas conforme demonstra o passo 2 da figura abaixo:

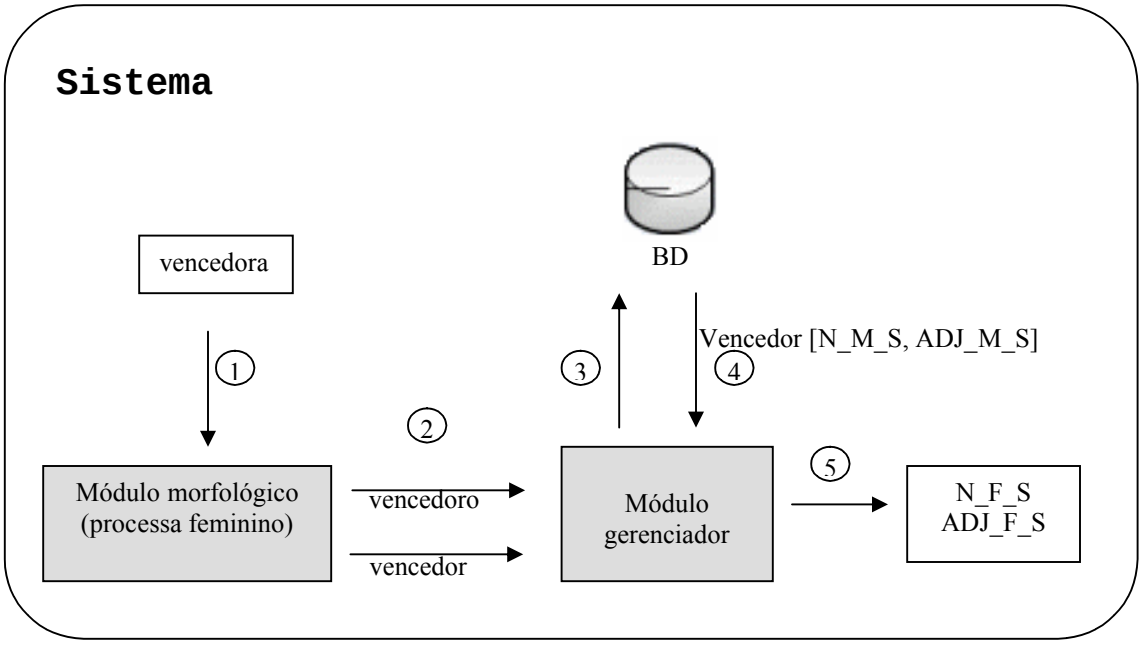

FIGURA 4.14 – PROCESSAMENTO DO GÊNERO

O esquema acima pode ser descrito da seguinte forma:

1 – Nesse passo, o **módulo gerenciador** já efetuou a busca pelo vocábulo "vencedora" no BD e nada encontrou. A partir daí, o vocábulo será enviado para o **módulo morfológico** que irá fazer as transformações necessárias.

2- O **módulo morfológico** efetua as transformações e apresenta duas saídas possíveis: "vencedoro" e "vencedor". Essas regras podem ser vistas nas linhas 2 e 3 da tabela 4.6. As duas saídas possíveis são, então, retornadas para o **módulo gerenciador**.

3- O **módulo gerenciador** efetua duas buscas no BD. Uma para "vencedoro" e outra para "vencedor".

4- Ao verificar ambos os vocábulos no BD, o **módulo gerenciador** só receberá retorno proveniente de “vencedor”, visto que “vencedoro” não faz parte do nosso vocabulário. As etiquetas N_M_S e ADJ_M_S serão modificadas dentro do **módulo gerenciador** para N_F_S e ADJ_F_S.

5- O módulo gerenciador apresenta as possíveis classes gramaticais (N_F_S e ADJ_F_S).

Cada terminação existente no feminino tem a sua contraparte que é uma terminação no masculino, conforme ilustra a tabela abaixo:

	Exemplo	**Feminino**	**Substitui**	**Masculino**	**Resultado**	
1.	aluna	-a	→	-o	aluno	Encontrado no BD
2.	autora	-ra	→	-r	autor	Encontrado no BD
3.	futura	-ra	→	-ro	futuro	Encontrado no BD
4.	patroa	-oa	→	-ão	patrão	Encontrado no BD

TABELA 4.6 – TRANSFORMAÇÃO DE GÊNERO FEMININO PARA MASCULINO

A tabela acima apresenta algumas terminações cadastradas no sistema. Para reduzir o número de buscas, algumas terminações mais especificas foram cadastradas, como é o caso de “ra”, “sa”, etc. A regra mais abrangente que foi cadastrada, representada na primeira linha na tabela acima, é a transformação de “a” para “o”. Existe uma outra regra mais abrangente que **não** foi cadastrada e é demonstrada na tabela abaixo:

Exemplo	**Feminino**	**substitui**	**Masculino**	**Resultado**	
Autora	-a	→	-	autor	Encontrado no BD

TABELA 4.7 – EXEMPLO DE REGRA DE TRANSFORMAÇÃO NÃO CADASTRADA

Se cadastrássemos a regra acima, veríamos, por exemplo, “autora” se transformando em “autor” e o sistema encontraria o resultado no BD

Entretanto, o sistema geraria mais de uma busca no BD para casos simples como “aluna”, conforme representa a tabela a seguir:

Exemplo	**Feminino**	**substitui**	**Masculino**	**Resultado**	
aluna	-a	→	-	alun	Não encontrado no BD
aluna	-a	→	-o	aluno	Encontrado no BD

TABELA 4.8 – EXEMPLO DE RESULTADO DE TRANSFORMAÇÃO PARA “ALUNA”

Observando a tabela acima, percebemos que o sistema irá buscar “alun” no BD e não encontrará nada. Embora “aluno” tenha sido encontrado, como demonstra a linha 2, o sistema fez duas buscas no BD, ou seja, levou o dobro do tempo que levaria para encontrar o vocábulo. Por essa razão, a regra da tabela 4.7 **não** foi cadastrada no sistema.

4.9.3 Grau dos substantivos e adjetivos

O processamento do grau também ocorre fundamentalmente a partir de regras. Um arquivo XML contém informações referentes aos sufixos de determinado grau relacionadas com informações de terminações nos vocábulos sem marcação de grau. Abaixo segue uma tabela representando um trecho do arquivo XML correspondente às regras referentes ao grau.

Forma alvo	Palavra	Sufixo de grau	Sufixo s/ modificação de grau	Grau
altíssimo	alto	-íssimo	o	superlativo
barcaça	barca	-aça	-	superlativo
homemzinho	homem	-zinho	-	diminutivo
bonitinho	bonito	-inho	o	diminutivo

TABELA 4.9 – EXEMPLO DE REGRAS DE GRAU

Todo o processamento ocorre baseado em transformações nos vocábulos recebidos. O processamento do módulo sintático existente para o grau ocorre de forma idêntica ao gênero. O sistema verifica a ocorrência de sufixos modificados (coluna 3 da tabela acima) em grau e, caso encontre, substitui os sufixos pelos seus correspondentes sufixos sem modificação de grau (coluna 4 da tabela acima).

Dessa forma, caso o sistema se depare com uma palavra no diminutivo, como por exemplo, “legalzinho”, fará uma busca no BD e não encontrará resultados. Sendo assim, irá acionar o módulo de processamento morfológico, que tem como uma de suas tarefas, a verificação dos morfemas cadastrados como formadores de diminutivos. Após essa verificação, o sistema irá retornar as possíveis realizações do vocábulo não modificado em grau, ou seja, no caso acima irá retornar o vocábulo desprovido do sufixo “zinho”, ou seja, “legal”. Com esse resultado o sistema irá verificar novamente a ocorrência no BD e, dessa vez, irá encontrar o vocábulo mencionado. O sistema, então, atribui as possíveis etiquetas existentes para “legal”, acrescidas da etiqueta “_DIMINUT”. Embora a etiqueta relacionada ao diminutivo não seja utilizada pelo sistema, foi acrescida para que pudesse auxiliar futuras melhorias. Ainda que possam existir formas menos produtivas no PB, como “numerinho”, foram consideradas apenas as formas de ocorrência mais significativas. O sistema também é capaz de lidar com formas como o diminutivo para “amigo” (amiguinho), substituindo a ocorrência “uinho” por “o” e, dessa forma, encontrando o vocábulo no BD.

4.10. Módulo de análise sintática

O grupo de *tokens* com suas informações gramaticais são enviados para o analisador sintático, que efetua o processamento de (n) regras gramaticais para desambiguizar os vocábulos que possuem mais de uma classificação gramatical, como demonstra o esquema abaixo:

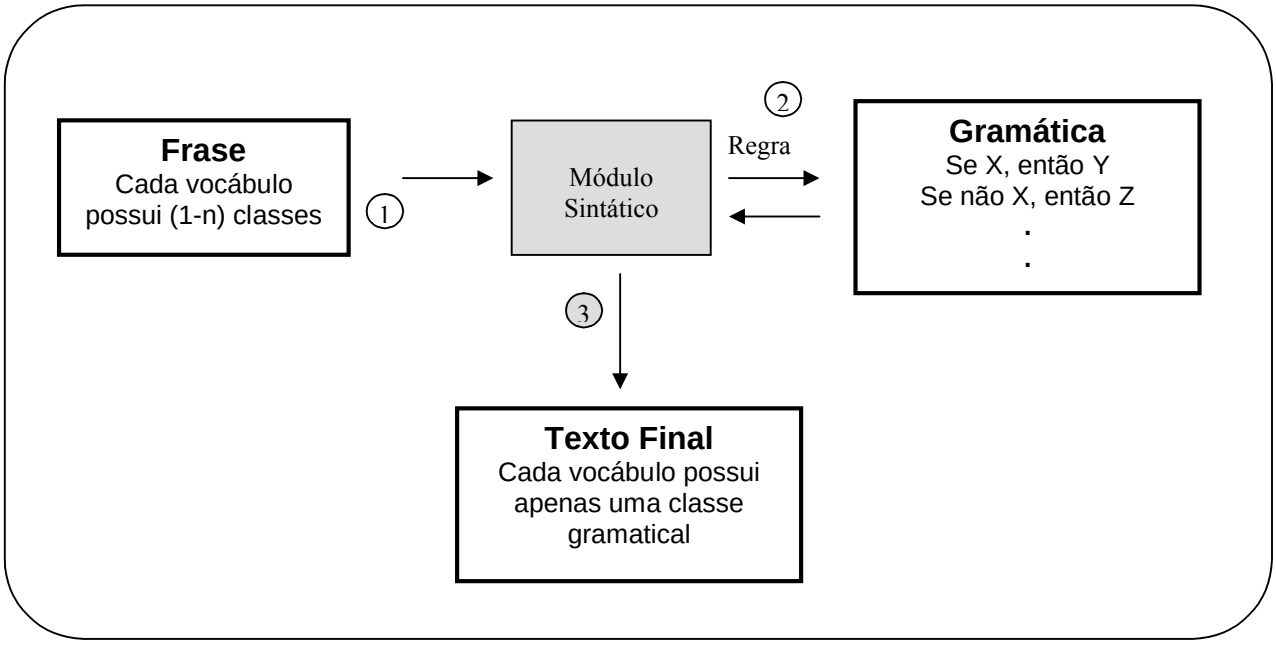

FIGURA 4.15 – MÓDULO DE ANÁLISE SINTÁTICA

O módulo gerenciador percorre, palavra por palavra, uma determinada frase. Caso haja, nessa frase, uma palavra com mais de uma classe gramatical, o sistema envia **toda** a frase para o módulo sintático. O esquema acima pode ser descrito da seguinte forma:

1 – Nesse passo, o **módulo sintático** recebe a frase (proveniente do **módulo gerenciador**) que contém (n) palavras com ambigüidade categorial. Todos os vocábulos já possuem suas respectivas etiquetas, para que o **módulo sintático** possa processar as regras gramaticais, baseando-se na classificação gramatical dos constituintes das frases.

2- O **módulo sintático** acessa as regras gramaticais para produzir um texto final, onde cada vocábulo possuirá apenas uma classe gramatical.

3- O **módulo sintático** apresenta o texto com os nomes e verbos identificados.

Embora o sistema classifique todos os constituintes da frase, só são apresentados os nomes e verbos, pois este é o objetivo do sistema: identificar nomes e verbos.

Para melhor entendimento do funcionamento do módulo sintático, apresentamos abaixo um algoritmo simplificado de como ocorre o processamento sintático interno.

```
x = 0;
leia posição(x);
enquanto posição(x) for diferente do marcador de fim de sentença:
        se posição(x) = substantivo,
                então leia posição(x-1);
                        se posição(x-1) = artigo,
                          então   posição(x) = substantivo;
        caso contrário, executa outras regras;
        incrementa(x);
        leia posição(x);
fim;
```

FIGURA 4.16 – ALGORITMO DE EXEMPLO

Para analisarmos o algoritmo acima, vamos utilizar um exemplo na prática:

(10) A casa é bonita.

Pressupondo que o algoritmo esteja processando a palavra "casa", a regra acima determina que se o vocábulo da posição atual ("casa") possuir a etiqueta de substantivo (é o caso de "casa") e se a posição anterior for um artigo (é o caso de "a"), o sistema irá atribuir a etiqueta de nome (N) ao vocábulo, **independente** do fato de "casa" se enquadrar na classificação de verbo também.

É importante ressaltar que o sistema procura etiquetar todos os vocábulos, independente de os vocábulos serem nomes e verbos. Caso o vocábulo seja encontrado no BD ou com alguma regra sintática, a etiqueta será atribuída ao mesmo. O sistema conta com 64 regras sintáticas. Se não etiquetássemos, por exemplo, os adjetivos, o sistema não poderia se utilizar da característica mais prototípica de precedência dos substantivos aos adjetivos. Caso o vocábulo seja encontrado no BD ou pela aplicação de alguma regra morfológica, a etiqueta será atribuída ao mesmo.

O maior problema encontrado na análise sintática foi convergência entre os adjetivos e as formas de particípio, visto que são casos de homonímia. O exemplo abaixo ilustra a ocorrência desse problema:

(11) "A série **exibida** aqui pela Cultura..."[15]

(12) "...e a minha melhor amiga começou a andar com uma menina **exibida**."[16]

---

[15] Retirado do Corpus de Extractus de Textos Eletrônicos NILC/ Folha de São Paulo.

[16] Referência encontrada no sistema de busca google™.

O vocábulo “exibida” pode ser particípio do verbo “exibir” ou adjetivo. Embora as análises lexicalistas tratem ambos como adjetivos (ver item 2.1.1), a metodologia utilizada mantém a diferença categorial, considerando verbos os particípios que só podem ser substituídos por verbos (11) e adjetivos os que podem também ser substituídos por adjetivos (12). Em função dessa decisão metodológica, o sistema apresentará algumas classificações incorretas. Para resolver esse tipo de ambigüidade categorial, seriam necessárias informações de caráter semântico e pragmático. A construção de um módulo que seja capaz de processar essas informações será objeto de trabalhos futuros.

Outro problema enfrentado foi a distinção entre adjetivos e substantivos. Como no português o substantivo costuma preceder o adjetivo e pelo fato de existirem palavras que podem ser classificadas como substantivo e adjetivo, como é o caso de “velho”, alguns cuidados tiveram que ser adotados. Por mais que a ordem vocabular não-marcada seja o substantivo preceder o adjetivo, sintagmas como “O velho maluco” e “O velho mundo” se tornam um problema. Quando buscados separadamente no BD, “velho” e “mundo” trarão as etiquetas conforme demonstra a figura abaixo:

```
vocábulo: velho
Classe(s): [ADJ_M_S, N_M_S]

vocábulo: mundo
Classe(s): [N_M_S]
```

FIGURA 4.17 – RESULTADO DE BUSCA PARA OS VOCÁBULOS “VELHO” E “MUNDO”

O vocábulo “velho” pode funcionar como substantivo ou adjetivo, mas “mundo" só pode ser substantivo. Em “o velho mundo”, “velho” só poderá ser adjetivo, visto que mundo será necessariamente substantivo. Apesar de o adjetivo estar precedendo o substantivo, o sistema será capaz de perceber essa precedência através das regras sintáticas e das etiquetas morfológicas associadas. Já em “o velho maluco”, como tanto “velho” como “maluco” possuem mais de uma classificação gramatical, a ordem mais prototípica será o fator determinante, ou seja, “velho” será classificado como “nome”. Por utilizar apenas informações morfossintáticas, o sistema irá convergir para o resultado com maior probabilidade de ocorrência, mas a idéia é que o sistema tenha módulos capazes de processar semântica e pragmática (processos discursivos) e dessa forma, seja capaz de processar informações contextuais, caso existam. O exemplo abaixo descreve de maneira simplificada a forma que o sistema poderia processar informações contextuais para convergir para o resultado com um menor índice de erros.

(13) *“E fora do ano novo tb, copacabana só tem velho, velho maluco, camelô, velho chato, camelô”* [17]

Podemos observar que na frase acima existe mais de uma ocorrência do vocábulo “velho”. Com um correto processamento de “tem velho”, o sistema concluiria que a primeira ocorrência seria substantivo. Ao fazer uma análise contextual, podemos perceber que em “velho maluco”, “velho” é o substantivo e “maluco” é o adjetivo. Da mesma forma, regras contextuais poderiam ser

[17] Referência encontrada no sistema de busca google™.

aplicadas pelo sistema para verificar que essa segunda ocorrência de “velho” também seria um substantivo.

Esse nível de comportamento pode ser estudado e adaptado ao sistema, mas não faz parte do escopo desta dissertação.

5. RESULTADOS OBTIDOS

Neste capítulo, serão apresentados os resultados obtidos para a identificação de *nomes* e *verbos* pelo "SOFIA". Para que seja possível visualizar os resultados práticos da implementação do sistema, a análise de dois textos que se enquadram em gêneros textuais diferentes será apresentada, com o objetivo de demonstrar o alto índice de acerto que o sistema obtém e, ao mesmo tempo, discutir os problemas que impediram uma taxa de acerto de 100%. Como ficará claro no decorrer da análise, tais problemas decorrem predominantemente da inexistência de um analisador semântico/pragmático no atual estágio do sistema.

Os textos são matérias jornalísticas retiradas do Jornal "Folha de São Paulo" (CETEN-Folha) e "O cravo e a rosa...", conto de Jorge Amado.

5.1. Texto jornalístico

Na análise dos dados foram processadas 116 frases do "Corpus de Textos Eletrônicos NILC/Folha de São Paulo(CETEN-Folha)", que totalizaram 1998 palavras. A saída obtida pelo sistema encontra-se no Apêndice C[18]. É importante ressaltar que alguns tratamentos foram

[18] Os nomes e verbos identificados a mais foram sublinhados, os nomes e verbos não identificados, ou seja, identificados a menos, foram marcados com negrito e os nomes que foram classificados como verbos assim como os verbos que foram classificados como nomes, se apresentarão sublinhados e em negrito. Todo vocábulo identificado a mais ou a menos irá apresentar entre colchetes a classificação correta.

efetuados no corpus para a extração de sentenças. Os títulos das reportagens e as citações não foram processados.

A tabela abaixo demonstra os resultados obtidos pelo sistema:

	Nomes	**Verbos**
Total (existente no corpus)	645	278
Identificados pelo sistema	654	288
Identificados corretamente	636	276
Identificados incorretamente (a mais)	18	12
Identificados incorretamente (a menos)	9	2

TABELA 4.10 – RESULTADOS OBTIDOS

Das 1998 palavras, 645 deveriam ser identificadas como *nomes* e 278 como *verbos*. Foram identificados 654 *nomes* pelo sistema. Desses *nomes*, 18 foram identificados a mais (incorretamente), ou seja, com a classe gramatical inadequada. Foram identificados a menos (não foram identificados) 9 nomes. Foram identificados 288 verbos pelo sistema. Desses *verbos*, 12 foram identificados incorretamente (a mais) e 2 foram identificados a menos (não foram identificados).

A eficácia dos métodos utilizados para a identificação de nomes e verbos foi medida em termos de precisão – percentual de nomes e verbos identificados que estavam corretos – e abrangência – percentual de nomes e verbos existentes no corpus de teste e que foram identificados corretamente.

A precisão que o sistema alcançou é demonstrada através da fórmula abaixo:

$$(5.1)\ \text{Precisão} = \frac{\text{Nomes e verbos identificados corretamente}}{\text{Total de nomes e verbos identificados}}$$

A fórmula acima é resolvida da seguinte forma:

$$(5.2)\ \text{Precisão} = \frac{\text{636 nomes + 276 verbos = 912 nomes e verbos}}{\text{654 nomes + 288 verbos = 942 nomes e verbos}}$$

A precisão alcançada foi de 96.8%.

A abrangência que o sistema alcançou é demonstrada a partir da seguinte fórmula:

$$(5.3)\ \text{Abrangência} = \frac{\text{Nomes e verbos identificados corretamente}}{\text{Total de nomes e verbos presentes no corpus}}$$

$$(5.4)\ \text{Abrangência} = \frac{\text{636 nomes + 276 verbos = 912 nomes e verbos}}{\text{645 nomes + 278 verbos = 923 nomes e verbos}}$$

A abrangência alcançada foi de 98.8%.

As fórmulas acima demonstram que o SOFIA apresenta índices de precisão e abrangência bastante satisfatórios. Na próxima seção, serão analisados os principais problemas de classificação encontrados, visto que a discussão desses casos pode contribuir para futuros refinamentos do sistema.

5.1.1 Problemas de classificação

A seguir, será feita uma descrição mais específica dos erros encontrados. O objetivo é avaliar os grupos de erros em questão.

Os 18 vocábulos identificados incorretamente (a mais) como *nome* se apresentam em 3 grupos distintos; 14 adjetivos classificados como nome, 3 advérbios classificados como nome e 1 verbo classificado como nome.

A tabela de abaixo apresenta esses resultados:

	Erro cometido	**Exemplo**	**Classificação obtida pelo sistema**	**Classificação correta**	**Quantidade de ocorrências**
1.	nome/adjetivo[19]	original	nome	adjetivo	14
2.	nome/advérbio[20]	Além	nome	advérbio	3
5.	nome/verbo[21]	descobertas	nome	verbo	1

TABELA 4.11 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MAIS

O maior problema encontrado com a classificação incorreta dos nomes envolve a ambigüidade entre *nomes* e *adjetivos.* A linha 1 da tabela acima demonstra que 14 *adjetivos* foram incorretamente classificados como *nomes*. Embora seja observada uma ambigüidade categorial entre *nome* e *adjetivo*, isso não significa que os vocábulos que sofreram a classificação incorreta

[19] Os casos podem ser encontrados nas frases 37, 40, 47, 69, 71, 74, 79, 83, 84, 90, 107, e 111 do Apêndice C.

[20] Os casos podem ser encontrados nas frases 45, 84 e 105 do Apêndice C.

[21] O casos pode ser encontrado na frase 77 do Apêndice C.

possuem apenas as duas classificações em questão (*nome/adjetivo*). Casos como "mesmo"[22] possuem ambigüidade categorial entre *nome*, *advérbio*, *pronome* e *adjetivo*, mas a análise não objetiva discutir as (n) possibilidades de classificação e sim mensurar e agrupar as classificações incorretas.

Como o problema mais freqüente recaiu sobre a distinção entre *nomes* e *adjetivos*, os principais tipos de erros que geraram essas classificações inadequadas serão discutidos a seguir.

O problema mais recorrente refere-se à necessidade de uma análise semântico/pragmática em determinados contextos. Por exemplo, o sistema apresentou dificuldade na identificação do *adjetivo* "preso" causada pela presença do verbo de ligação:

(14) "Os três estão presos..." (*AC*[23] *37*)

Esse mesmo problema ocorre na classificação do *adjetivo* "superiores", como demonstra o exemplo abaixo:

(15) "...se os escravos fossem superiores..." (*AC 71*)

Contrastemos o exemplo (15) com o caso a seguir:

(16) "Se os mariscos fossem conchas..."

---

[22] Caso encontrado na frase 47 do Apêndice C

Enquanto no exemplo acima, o vocábulo “conchas” deve ser classificado como *substantivo*, em contexto semelhante o vocábulo “superiores” no exemplo (15) deve ser classificado como *adjetivo*. A ambigüidade entre *nome* e *adjetivo* em contextos subseqüentes a verbos de ligação acabaram gerando 5 erros. Problemas como esses só poderiam ser resolvidos com uma análise semântico/pragmática.

No exemplo a seguir, o vocábulo “menor” ocorre relacionado a um elemento citado no início da frase (valor). Esse problema também pode ser resolvido com a análise do contexto, que será feita em trabalhos futuros:

(17) “O valor é mais que o dobro do estimado pela Exxon, mas menor que o original...” (*AC 83*)

Outra questão que merece destaque diz respeito às dificuldades impostas pelas *conjunções coordenativas*. Embora o sistema possua regras para tratamento de *conjunções*, houve problemas relacionados às *conjunções coordenativas* “e” e “ou” na desambiguação de *nomes* e *adjetivos*, como ilustram os exemplos a seguir:

(20) “...o mesmo **ou** maior destaque...” (*AC 47*)

(21) “...leis complementares **e** ordinárias...” (*AC 90*)

---

[23] Os Apêndices serão representados da seguinte forma: Apêndice B será representado por AB e Apêncide C por AC e os números subsequentes indicam as linhas no caso do Apêndice B e o exemplo no caso do Apêndice C.

Os vocábulos “mesmo” e “ordinárias” foram identificados incorretamente como *nomes* porque o sistema ainda não é capaz de distinguir se as *conjunções coordenativas* unem dois termos ou duas orações, como seria o caso dos exemplos abaixo:

(22) Os jovens gostam de futebol e vôlei.

(23) As pessoas jogam futebol e vôlei não.

Nos exemplos acima, “futebol” e “vôlei” podem ser termos coordenados de um sintagma (22) ou argumentos de orações coordenadas (23), o que demandaria um tratamento sintático mais refinado do sistema.

O sistema identificou 9 *nomes* a menos, ou seja, identificou os vocábulos que deveriam ser considerados *nome* com outra classificação. Abaixo segue uma tabela apresentando um resumo do problema:

	Erro cometido	**Exemplo**	**Classificação obtida pelo sistema**	**Classificação correta**	**Quantidade de ocorrências**
1.	nome e adjetivo[24]	petroleiro	adjetivo	nome	4
2.	nome e verbo[25]	decorrer	verbo	nome	5

TABELA 4.12 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MENOS

O SOFIA identifica os vocábulos “petroleiro”, “delegado” e “estudante” incorretamente como *adjetivo*:

[24] Casos encontrados nas frases 81, 84, 114 e 115 do Apêndice C

(24) “O derramamento de óleo do petroleiro Exxon....” (*AC 81*)

(25) “...a estudante Claudirene Contijo....” (*AC 114*)

(26) “O delegado Celso Amorin...” (*AC 115*)

Isso ocorre porque o sistema encontrou um vocábulo sucessor que só poderia ser classificado como *nome*. Não existem regras capazes de verificar a ocorrência de um vocábulo sucessor com letra maiúscula, mas o sistema será adequado em projetos futuros, podendo então, corrigir esse erro gerado.

Com relação aos *verbos*, o sistema apresentou 12 *verbos* identificados a mais, ou seja, os vocábulos possuíam outra classe e foram classificados como *verbo*. A tabela a seguir resume os resultados:

	Erro cometido	**Exemplo**	**Classificação obtida pelo sistema**	**Classificação correta**	**Quantidade de ocorrências**
1.	verbo/adjetivo[26]	original	verbo	adjetivo	6
2.	verbo/nome[27]	parte	verbo	nome	5
5.	verbo/conjunção[28]	como	verbo	conjunção	1

TABELA 4.13 – RESUMO DA CLASSIFICAÇÃO DE VERBOS A MAIS

[25] Casos encontrados nas frases 18, 42, 69, 88 e 111 do Apêndice C

[26] Os casos podem ser encontrados nas frases 31, 62, 71, 76, 105 e 112 do Apêndice C.

[27] Os casos podem ser encontrados nas frases 17, 42, 69, 88 e 111 do Apêndice C.

[28] O caso pode ser encontrado na frase 89 do Apêndice C.

O maior número de erros gerados ocorreu em *adjetivos* (6 ocorrências) que foram classificados como *verbos*, em que todos esses *verbos* possuíam ocorrência no particípio. Conforme já descrito anteriormente, a metodologia utilizada para esse problema envolve a possibilidade de substituição do *verbo/adjetivo* por um *adjetivo* que não seja um *verbo* no particípio. Caso possa ser substituído por tal *adjetivo*, o vocábulo será considerado *adjetivo*; caso contrário será considerado *verbo*. Visto que essa categoria é problemática e discutida por diversos autores, essa foi a solução encontrada que pareceu mais prática em relação a desdobramentos futuros do sistema. Analisemos o trecho abaixo:

(27) "...a justificativa moral para mantê-los escravizados." (*AC 70*)

Após analisar a frase acima, chegou-se à conclusão de que o sistema identificou o vocábulo "escravizados" incorretamente. De acordo com os critérios adotados neste trabalho, o vocábulo teria que ser classificado como *adjetivo* pelo fato de poder ser substituído por um *adjetivo* que não seja um particípio, como, por exemplo, "alegres". O sistema o classificou como *verbo* por ter considerado que o item fazia parte de uma perífrase verbal.

Houve 5 casos em que *nomes* foram incorretamente classificados como *verbos*. O exemplo abaixo descreve o problema encontrado:

(28) "Parte dos recursos para a formação do FSE..." (*AC 17*)

O vocábulo "parte" foi incorretamente classificado como *verbo*. Trata-se de um exemplo que só poderia ser resolvido com análises semântica e pragmática, visto que o contexto sintático admite ocorrências em que "parte" seria classificado como *verbo*:

(29) Parte do mais fraco o pedido de desistência.

Com base nos exemplos 28 e 29, constatamos que será o contexto que determinará a classificação de ”parte” como *verbo* ou como *nome*, visto que a sintaxe não consegue dar conta nesses casos.

Com relação à classificação incorreta de *verbos* (a menos) a tabela abaixo indica as ocorrências:

	Erro cometido	**Exemplo**	**Classificação obtida pelo sistema**	**Classificação correta**	**Quantidade de ocorrências**
1.	vebo e adjetivo[29]	exposta	adjetivo	verbo	1
2.	verbo e nome[30]	descobertas	nome	verbo	1

TABELA 4.14 – RESUMO DA CLASSIFICAÇÃO DE VERBOS A MENOS

O vocábulo “exposta” (*AC 73*) e o vocábulo “descobertas” (*AC 78*) apresentam ambigüidade envolvendo *verbos* no particípio. Os erros em questão também são decorrentes da ausência de um analisador semântico/pragmático.

O sistema utilizou regras sintáticas em 26% dos nomes e 17% dos verbos presentes nas frases, pois os mesmos possuíam ambigüidade categorial. Os 74% de nomes e 83% dos verbos restantes não possuíam ambigüidade.

[29] Caso encontrado na frase 73 do Apêndice C

[30] Caso encontrado na frase 78 do Apêndice C

Conforme mencionado, o sistema utilizou regras sintáticas em 531 vocábulos, dos quais 168 eram *nomes* e 47 eram *verbos*. Embora o sistema tenha aplicado regras de desambiguação em aproximadamente 1/4 dos nomes e 1/6 dos verbos, tais regras foram utilizadas em vocábulos com ambigüidades que não possuíam classificação gramatical de *nome* ou *verbo* (ex: ambigüidade entre *adjetivo* e *advérbio*). Os gráficos abaixo apresentam a distribuição entre *nomes* e *verbos*, destacando a ambigüidade e os erros encontrados pelo sistema.

A figura 4.18 destaca a porcentagem dos *nomes* identificados corretamente e dos *nomes* identificados incorretamente (a menos), visto que os *nomes* identificados incorretamente (a mais) não fazem parte do número total de *nomes* existentes no texto.

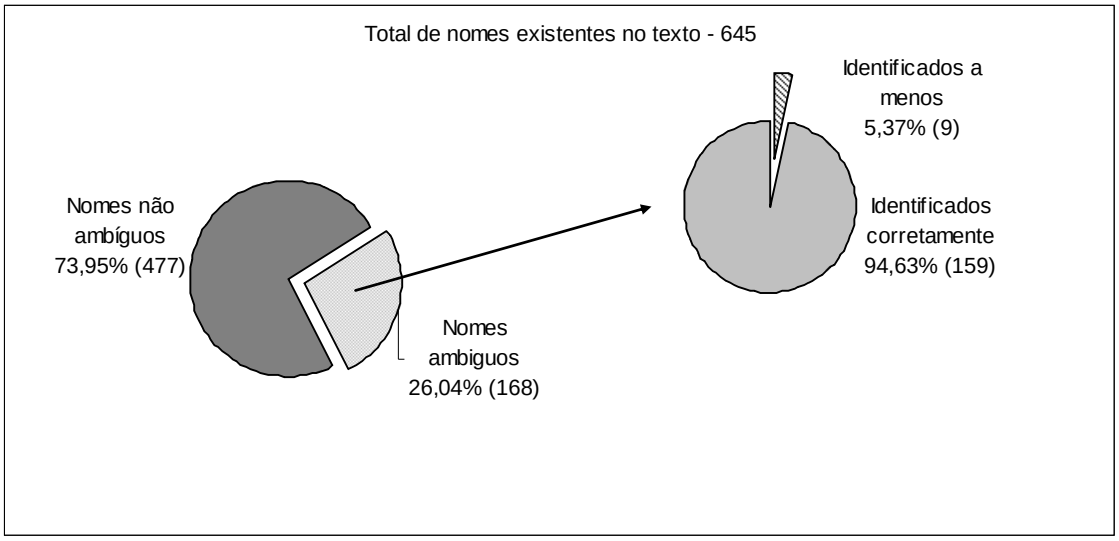

FIGURA 4.18 – USO DE REGRAS EM NOMES AMBÍGUOS

A figura 4.19 destaca a porcentagem dos *verbos* identificados corretamente e dos *verbos* identificados incorretamente (a menos). Da mesma forma que na figura anterior, os *verbos*

identificados incorretamente (a mais) não fazem parte do número total de *verbos* existentes no texto.

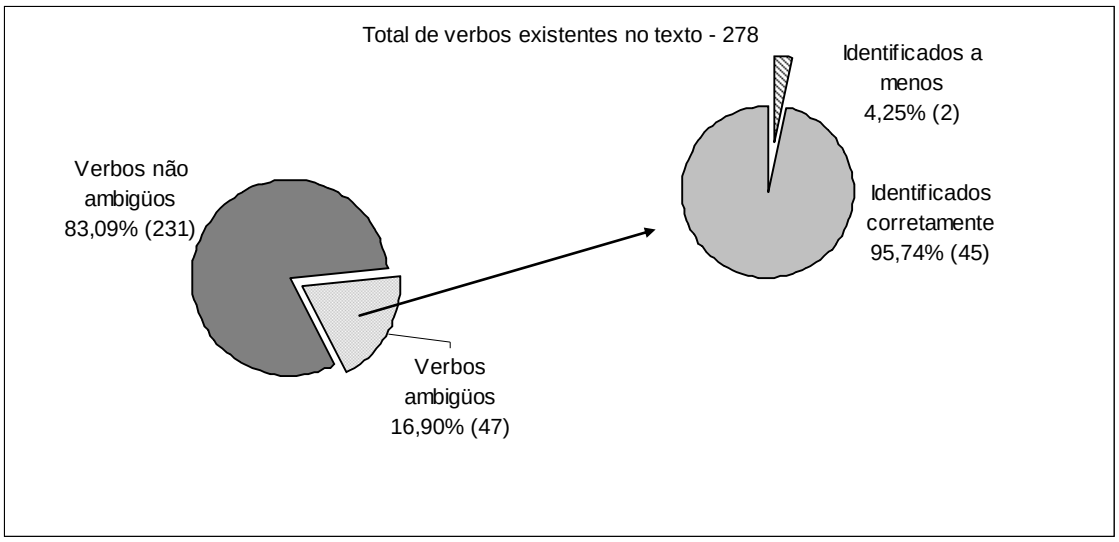

FIGURA 4.19 – USO DE REGRAS EM VERBOS AMBÍGUOS

Os gráficos acima ratificam que o sistema obteve bons índices de acerto para a identificação de nomes e verbos.

5.2. Texto literário

Para nos certificarmos de que SOFIA é capaz de lidar com textos não-jornalísticos, inserimos um texto literário, intitulado “Nem a rosa, nem o cravo...” de Jorge Amado. A saída apresentada pelo sistema encontra-se no apêndice B com os devidos *nomes* e *verbos* identificados.

O sistema apresentou bons resultados, confirmando a hipótese de que é capaz de lidar com textos de contextos distintos, conforme ilustra a tabela abaixo:

	Nomes	**Verbos**
Total (existente no corpus)	214	98
Identificados pelo sistema	216	102
Identificados corretamente	210	98
Identificados incorretamente (a mais)	6	4
Identificados incorretamente (a menos)	4	0

TABELA 4.15 – RESULTADOS OBTIDOS NO TEXTO LITERÁRIO

O texto possui 774 palavras, dentre as quais 214 são *nomes* e 98 são verbos. O sistema apresentou 216 *nomes* identificados, entretanto, identificou 6 *nomes* incorretamente a mais e 4 *nomes* incorretamente a menos. Além disso, identificou 102 *verbos*, sendo que nenhum deles foi identificado incorretamente a menos; houve apenas 4 casos de identificações incorretas a mais.

A eficácia do reconhecimento de nomes e verbos no texto literário também será apresentada em termos de precisão e abrangência.

Através da substituição na fórmula 5.1, chegamos à precisão alcançada pelo sistema para o texto em questão.

$$(5.5)\quad \text{Precisão} = \frac{\text{210 nomes + 98 verbos = 308 nomes e verbos}}{\text{216 nomes + 102 verbos = 318 nomes e verbos}}$$

O sistema atingiu uma precisão de 96.8% de acerto. Abaixo podemos visualizar o resultado para a abrangência conforme a fórmula 5.3:

$$(5.6)\quad \text{Abrangência} = \frac{308 \text{ nomes e verbos}}{214 \text{ nomes} + 98 \text{ verbos} = 312 \text{ nomes e verbos}}$$

A abrangência alcançada foi de 98.7%.

Os índices acima demonstram que o sistema também se mostrou altamente eficaz para o gênero literário. A exemplo do que foi feito para o texto anterior, serão discutidas as dificuldades encontradas pelo sistema para o texto selecionado.

5.2.1 Problemas de classificação

As tabelas subsequentes irão apresentar informações mais detalhadas sobre os erros ocorridos. Primeiramente, na tabela a seguir, podemos observar os detalhes sobre os 6 *nomes* que foram identificados, incorretamente, a mais:

	Erro cometido	**Exemplo**	**Classificação obtida pelo sistema**	**Classificação correta**	**Quantidade de ocorrências**
1.	nome e adjetivo[31]	loiro	nome	adjetivo	4
2.	nome e advérbio[32]	hoje	nome	advérbio	2

TABELA 4.16 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MAIS

[31] Casos encontrados nas linhas 6, 15, 24, 37 do Apêndice B

[32] Casos encontrados na linha 36 (“sempre” e “ontem”) do Apêndice B

As ocorrências que geraram erros de identificação de nomes incorretamente (a mais) foram causadas por vocábulos com ambigüidade entre *nome* e *adjetivo* e ambigüidade entre *nome* e *advérbio*. A ausência de análise semântico/pragmática ocasionou todos esses erros, como ocorreu em casos semelhantes no texto jornalístico.

Abaixo segue um exemplo da ambigüidade entre *nome* e *adjetivo*:

(30) "Já viste um loiro **trigal** balançando ao vento?" (*AB 6*)

O sistema identificou "loiro" como *nome*, embora loiro seja *adjetivo* e "trigal" seja o *nome* nesse contexto. Visto que tanto "loiro" como "trigal" podem ser adjetivos, o sistema atribuiu a classe *nome* ao vocábulo "loiro" pelo fato de ter optado pela forma mais prototípica de o *nome* preceder o *adjetivo*.

Na linha 2 da tabela 4.16, vemos a ocorrência de dois casos em que o sistema identificou um *nome* como *advérbio*. Abaixo segue o trecho retirado do texto:

(31) "... com as palavras de sempre, com as frases de ontem..." (*AB 36*)

A identificação incorreta dos *advérbios* acima ocorreu pelo fato de existir a preposição "de" antes dos *advérbios*. Há outros casos como "hoje", em que o sistema classifica corretamente por não existir preposição antes do vocábulo, como segue abaixo:

(32) "Hoje só o ódio pode fazer com que o amor perdure sobre o mundo." (*AB 48*)

A tabela abaixo apresenta os resultados obtidos para a classificação incorreta dos *nomes* (a menos).

	Erro cometido	Exemplo	Classificação obtida pelo sistema	Classificação correta	Quantidade de ocorrências
1.	nome e adjetivo[33]	trigal	adjetivo	nome	2
2.	nome e advérbio[34]	tardes	advérbio	nome	1
3.	nome e verbo[35]	madrugadas	verbo	nome	1

TABELA 4.17 – RESUMO DA CLASSIFICAÇÃO DE NOMES A MENOS

Podemos perceber que dois *nomes* foram identificados como *adjetivos*:

(33) "...um loiro trigal balançando..." (*AB 6*)

(34) "...ao passo das bestas hitleristas..." (*AB 24*)

O vocábulo "trigal" deveria ter sido classificado como *nome*, mas conforme foi descrito, a ordem mais prototípica prevaleceu, já que "loiro", também ambíguo, foi classificado como *nome*. De forma análoga, em 34, a ambigüidade entre *nome* e *adjetivo* existente em "bestas" e em "hitleristas" evidencia a dupla possibilidade de interpretação. Mais uma vez, verificamos que a resolução de problemas dessa natureza requereria um tratamento de ordem semântico/pragmática.

[33] Casos encontrados nas linhas 6 e 24 do Apêndice B

[34] Caso encontrado na linha 15 do Apêndice B

[35] Casos encontrados na linhas 26 e 5 do Apêndice B

A classificação verbal também gerou alguns erros. Quatro *verbos* foram identificados a mais, ou seja, incorretamente. A tabela abaixo representa cada um dos erros:

	Erro cometido	Exemplo	Classificação obtida pelo sistema	Classificação correta	Quantidade de ocorrências
1.	verbo e adjetivo[36]	inesperada	verbo	adjetivo	2
2.	verbo e advérbio[37]	junto	verbo	advérbio	1
3.	verbo e nome[38]	madrugadas	verbo	nome	1

TABELA 4.18 – RESUMO DA CLASSIFICAÇÃO DE VERBOS A MAIS

Na linha 1, observamos um caso em que o vocábulo "inesperada" pode ser *adjetivo e verbo*. Abaixo segue o exemplo extraído do texto:

(35) "...quando as bestas soltas no mundo..." (*AB 5*)

(36) "É como uma nuvem inesperada num céu azul e límpido." (*AB 26*)

O vocábulo "inesperada" pode ser substituído por outro *adjetivo*, como, por exemplo, "bonita", e, portanto, foi considerado *adjetivo*. Da mesma forma, ocorre com o vocábulo "soltas".

O sistema, entretanto, classificou esses vocábulos como *verbo* por não possuir ainda regras semântico/pragmáticas para fazer esse tipo de desambiguação.

[36] Caso encontrado na linha 5 e 26 do Apêndice B

[37] Caso encontrados na linha 47 do Apêndice B

[38] Caso encontrado na linha 22 do Apêndice B

No texto em questão, o sistema utilizou regras de desambiguação em 249 vocábulos, dentre os quais 76 nomes e 16 verbos utilizaram regras de desambiguação. Foram identificados 4 verbos a mais pelo sistema. Abaixo seguem duas figuras que demonstram mais claramente os resultados:

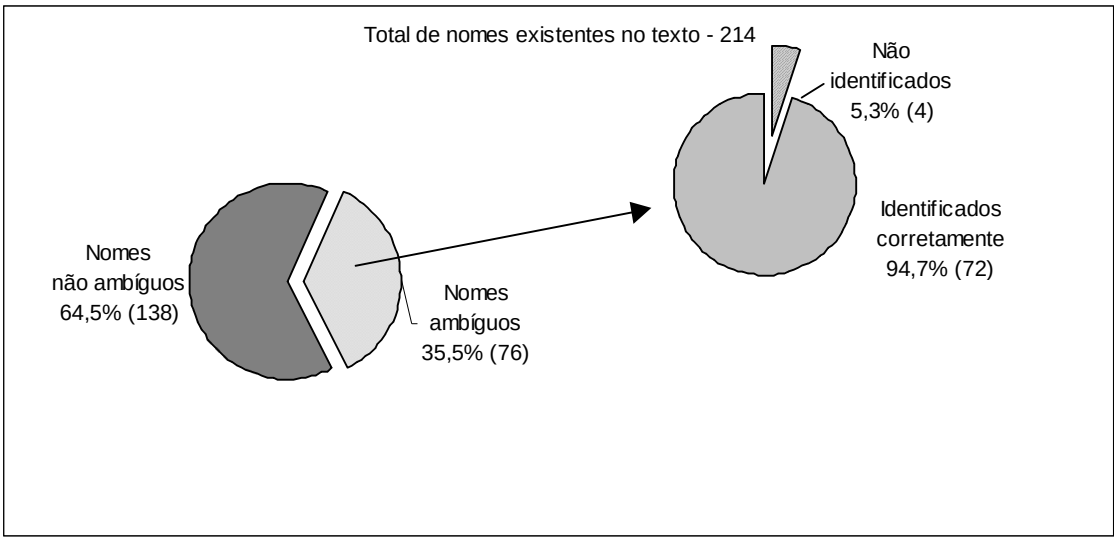

FIGURA 4.20 – USO DE REGRAS EM NOMES AMBÍGUOS

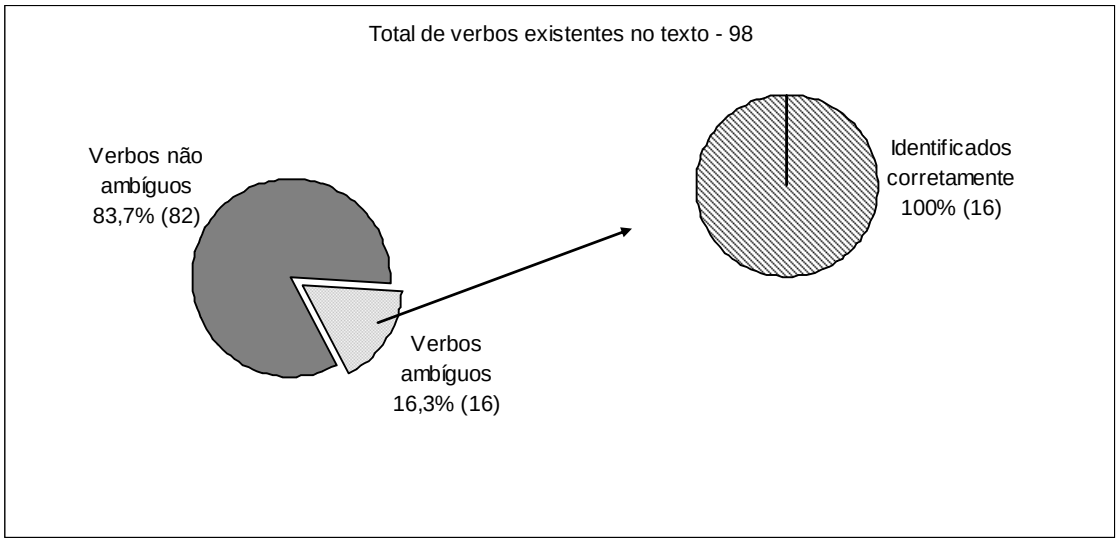

FIGURA 4.21 – USO DE REGRAS EM VERBOS AMBÍGUOS

Os nomes ambíguos obtiveram um índice de 94,7% de acerto. Na figura 4.21 podemos ver que 100% dos verbos ambíguos foram identificados corretamente.

Conforme mencionado, os resultados foram obtidos utilizando-se um número pequeno de regras com grande abrangência.

6. CONCLUSÃO

Este trabalho objetivou a construção de um analisador morfossintático capaz de reconhecer nomes e verbos para o Português do Brasil. O sistema possui atualmente 660 métodos em suas 81 classes desenvolvidas. A proposta foi estabelecer a criação de regras que fossem satisfatórias para resolução dos casos mais comuns para o PB. Os índices de 96,8% de precisão e 98,8% de abrangência para o texto jornalístico e os índices de 96,8% de precisão e 98,7% de abrangência para o texto literário podem ser considerados excelentes, visto que, muitas vezes, a desambiguação dos pares homônimos é complexa, até mesmo, para os falantes do PB.

É importante ressaltar que nenhum dos tratamentos pode ser dado como definitivo. Existem diversas melhorias que podem ser efetuadas no SOFIA. Como não foram utilizadas informações de concordância, não foram desenvolvidas regras de inter-relação entre os elementos considerando, por exemplo, o gênero dos elementos envolvidos. Também foram encontrados problemas relativos às palavras acentuadas que possuem ocorrência de número, aumentativo ou diminutivo, como é o caso de "número". Ao reconhecer o sufixo "zinho" em "numerozinho", o sistema irá efetuar o processamento morfológico para recuperar a forma não modificada do vocábulo supracitado. Nesse momento, irá derivar para o resultado "numero", não acentuado. Essa ocorrência não irá ser encontrada no BD, pois "número" é cadastrado com acentuação. Além disso, seria interessante criar um módulo acoplado ao banco de dados verbal capaz de reconhecer verbos não cadastrados no sistema.

No atual estágio do SOFIA, o principal problema encontrado está relacionado à inexistência de um analisador semântico/pragmático. O sistema, entretanto, está preparado para a futura

inclusão de novos módulos, em virtude de sua arquitetura flexível, que reflete a hipótese inicial de que os aspectos semânticos e pragmáticos exercem um papel fundamental na língua.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Barbosa, F.; Ferrari, L e Resende Jr, F.G. A distinçao entrre homógrafos heterófonos em Sistemas de Conversão texto-fala. In Silva, A.S; Torres, A; Gonçalves, M. Linguagem, Cultura e Cognição: Estudos em Lingüística Cognitiva. Coimbra: Almedina, 2004.

Barros, F. A. e Robin, J. Processamento de linguagem natural. Tutorial apresentado na Jornada de Atualização em Informática no Congresso da Sociedade Brasileira de Computação, 1997.

Basílio, M. M. P. Formação e classes de palavras no Português do Brasil. 1. ed. São Paulo: Contexto, 2004.

Câmara Jr., J. M. Dicionário de lingüística e gramática 9 ed. Petrópolis: Vozes, 1981.

Guedes, G.P.; Ferrari, L. Pares de homógrafos heterófonos em sistemas de conversão texto-fala. Rio de Janeiro:UFRJ. Mimeo, 2004.

Langacker, R. Foundations of cognitive grammar. Stanford: Stanford University Press. 1987.

Lemle, M. Análise Sintática. São Paulo: Ática. 1984

Lima, C. H. R. Gramática da língua portuguesa. 44ª ed. Rio de Janeiro: José Olympio, 2005.

Manning, C; Schütze, H. Foundations of statistical natural language processing, Cambridge, MA: MIT Press, 1999.

Nestrovski, A. Figuras do Brasil: 80 autores em 80 anos de Folha. São Paulo: PubliFolha, 2001.

Pimenta-Bueno, M. Os particípios passivos portugueses: verbos, adjetivos ou uma terceira classe? Comunicação ao VI Encontro Nacional de Lingüística. Rio de Janeiro, PVC, 1981.

APÊNDICE A – RÓTULOS UTILIZADOS NO SISTEMA

Rótulo	Descrição
M	Masculino
F	Feminino
I	Invariável

TABELA A-1 – RÓTULOS UTILIZADOS PARA IDENTIFICAR O GÊNERO

Rótulo	Descrição
S	Singular
P	Plural
I	Invariável

TABELA A-2 – RÓTULOS UTILIZADOS PARA IDENTIFICAR O NÚMERO

Rótulo	Descrição
FUT_PRES_IND	Futuro do presente do Indicativo
FUT_PRET_IND	Futuro do pretérito do Indicativo
PRES_IND	Presente do Indicativo
IMP_IND	Pretérito Imperfeito do Indicativo
PERF_IND	Pretérito Perfeito do Indicativo
MAIS_Q_PERF_IND	Pretérito-Mais-que-Perfeito
INF_PESS	Infinitivo Pessoal
V_FUT_SUBJ	Futuro do Subjuntivo
PRES_SUBJ	Presente do Subjuntivo
IMP_SUB	Pretérito Imperfeito do Subjuntivo
PART	Particípio Passado
AFR_IMP	Afirmativo Imperativo
NEG_IMP	Negativo Imperativo
GER	Gerúndio

TABELA A-3 – RÓTULOS UTILIZADOS PARA IDENTIFICAR O NÚMERO

Rótulo	Descrição
AUMENT	Aumentativo
DIMINUT	Diminutivo

TABELA A-4 – RÓTULOS UTILIZADOS PARA IDENTIFICAR O NÚMERO

Rótulo	Descrição
1	Primeira pessoa
2	Segunda Pessoa
3	Terceira Pessoa

TABELA A-5 – RÓTULOS UTILIZADOS PARA IDENTIFICAR AS PESSOAS VERBAIS

Rótulo	Descrição
DEF	Definido
INDEF	Indefinido

TABELA A-6 – RÓTULOS UTILIZADOS PARA IDENTIFICAR OS ARTIGOS

APÊNDICE B – NEM A ROSA NEM O CRAVO... (JORGE AMADO)

As frases(N) perdem(V) seu sentido(N), as palavras(N) perdem(V) sua significação(N) costumeira, como dizer(V) das árvores(N) e das flores(N), dos teus olhos(N) e do mar(N), das canoas(N) e do cais(N), das borboletas(N) nas árvores(N), quando as crianças(N) são(V) assassinadas(V) friamente pelos nazistas(N)? Como falar(V) da gratuita beleza(N) dos campos(N) e das cidades(N), quando as bestas(N) soltas(V)*[ADJ]* no mundo(N) ainda destroem(V) os campos(N) e as cidades(N)? Já viste(V) um loiro(N)*[ADJ]* **trigal***[N]* balançando(V) ao vento(N)? É(V) das coisas(N) mais belas do mundo(N), mas os hitleristas(N) e seus cães(N) danados destruíram(V) os trigais(N) e os povos(N) morrem(V) de fome(N). Como falar(V), então, da beleza(N), dessa beleza(N) simples e pura da farinha(N) e do pão(N), da água(N) da fonte(N), do céu(N) azul, do teu rosto(N) na tarde(N)? Não posso(V) falar(V) dessas coisas(N) de todos os dias(N), dessas alegrias(N) de todos os instantes(N). Porque elas estão(V) perigando(V), todas elas, os trigais(N) e o pão(N), a farinha(N) e a água(N), o céu(N), o mar(N) e teu rosto(N). Contra tudo que é(V) a beleza(N) cotidiana do homem(N), o nazifascismo(N) se levantou(V), monstro(N) medieval de torpe visão(N), de ávido apetite(N) assassino. Outros que falem(V), se quiserem(V), das árvores(N) nas **tardes***[N]* agrestes(N)*[ADJ]*, das rosas(N) em coloridos(N) variados, das flores(N) simples e dos versos(N) mais belos e mais tristes. Outros que falem(V) as grandes palavras(N) de amor(N) para a bem-amada(N), outros que digam(V) dos crepúsculos(N) e das noites(N) de estrelas(N). Não tenho(V) palavras(N), não tenho(V) frases(N), vejo(V) as árvores(N), os pássaros(N) e a tarde(N), vejo(V) teus olhos(N), vejo(V) o crepúsculo(N) bordando(V) a cidade(N). Mas sobre todos esses quadros(N) bóiam(V) cadáveres(N) de crianças(N) que os nazis(N) mataram(V), ao canto(N) dos pássaros(N) se mesclam(V) os gritos(N) dos velhos torturados(N) nos campos(N) de concentração(N), nos crepúsculos(N) se fundem(V) madrugadas(V)*[N]* de reféns(N) fuzilados. E, quando a paisagem(N) lembra(V) o campo(N), o que eu vejo(V) são(V) os trigais(N) destruídos(V) ao passo(N) das **bestas***[N]* hitleristas(N)*[ADJ]*, os trigais(N) que alimentavam(V) antes as populações(N) livres. Sobre toda a beleza(N) paira(V) a sombra(N) da escravidão(N). É(V) como uma nuvem(N) inesperada(V)*[ADJ]* num céu(N) azul e límpido. Como então encontrar(V) palavras(N) inocentes, doces palavras(N) cariciosas, versos(N) suaves e tristes? Perdi(V) o sentido(N) destas palavras(N), destas frases(N), elas me soam(V) como uma traição(N) neste momento(N). Mas sei(V) todas as palavras(N) de ódio(N), do ódio(N) mais profundo e mais mortal. Eles matam(V) crianças(N) e essa é(V) a sua maneira(N) de brincar(V) o mais inocente dos

brinquedos(N). Eles desonram(V) a beleza(N) das mulheres(N) nos leitos(N) imundos e essa é(V) a sua maneira(N) mais romântica de amar(V). Eles torturam(V) os homens(N) nos campos(N) de concentração(N) e essa é(V) a sua maneira(N) mais simples de construir(V) o mundo(N). Eles invadiram(V) as pátrias(N), escravizaram(V) os povos(N), e esse é(V) o ideal(N) que levam(V) no coração(N) de lama(N). Como então ficar(V) de olhos(N) fechados(V) para tudo isto e falar(V), com as palavras(N) de sempre(N)*[ADV]*, com as frases(N) de ontem(N)*[ADV]*, sobre a paisagem(N) e os pássaros(N), a tarde(N) e os teus olhos(N)? É(V) impossível(N)*[ADJ]* porque os monstros(N) estão(V) sobre o mundo(N) soltos e vorazes, a boca(N) escorrendo(V) sangue(N), os olhos(N) amarelos, na ambição(N) de escravizar(V). Os monstros(N) pardos, os monstros(N) negros e os monstros(N) verdes. Mas eu sei(V) todas as palavras(N) de ódio(N) e essas, sim, têm(V) um significado(N) neste momento(N). Houve(V) um dia(N) em que eu falei(V) do amor(N) e encontrei(V) para ele os mais doces vocábulos(N), as frases(N) mais trabalhadas. Hoje só o ódio(N) pode(V) fazer(V) com que o amor(N) perdure(V) sobre o mundo(N). Só o ódio(N) ao fascismo(N), mas um ódio(N) mortal, um ódio(N) sem perdão(N), um ódio(N) que venha(V) do coração(N) e que nos tome(V) todo, que se faça(V) dono(N) de todas as nossas palavras(N), que nos impeça(V) de ver(V) qualquer espetáculo(N) - desde o crepúsculo(N) aos olhos(N) da amada(N) - sem que junto(V)*[ADV]* a ele vejamos(V) o perigo(N) que os cerca(V). Jamais as tardes(N) seriam(V) doces e jamais as madrugadas(N) seriam(V) de esperança(N). Jamais os livros(N) diriam(V) coisas(N) belas, nunca mais seria(V) escrito(V) um verso(N) de amor(N). Sobre toda a beleza(N) do mundo(N), sobre a farinha(N) e o pão(N), sobre a pura água(N) da fonte(N) e sobre o mar(N), sobre teus olhos(N) também, se debruçaria(V) a desonra(N) que é(V) o nazifascismo(N), se eles tivessem(V) conseguido(V) dominar(V) o mundo(N). Não restaria(V) nenhuma parcela(N) de beleza(N), a mais mínima. Amanhã saberei(V) de novo palavras(N) doces e frases(N) cariciosas. Hoje só sei(V) palavras(N) de ódio(N), palavras(N) de morte(N). Não encontrarás(V) um cravo(N) ou uma rosa(N), uma flor(N) na minha literatura(N). Mas encontrarás(V) um punhal(N) ou um fuzil(N), encontrarás(V) uma arma(N) contra os inimigos(N) da beleza(N), contra aqueles que amam(V) as trevas(N) e a desgraça(N), a lama(N) e os esgotos(N), contra esses restos(N) de podridão(N) que sonharam(V) esmagar(V) a poesia(N), o amor(N) e a liberdade(N)!

APÊNDICE C – CETEM FOLHA

1. Muito mais do que nos tempos(N) na ditadura(N), a solidez(N) do PT(N) está(V), agora, ameaçada.
2. Nem Lula(N) nem o partido(N) ainda encontraram(V) um discurso(N) para se diferenciar(V).
3. Eles se dizem(V) oposição(N), mas ainda não informaram(V) o que vão(V) combater(V).
4. Muitas das prioridades(N) do novo governo(N) coincidem(V) com as prioridades(N) do PT(N).
5. A série(N) exibida(V) aqui pela Cultura(N) estreou(V) na TVI(N) de Portugal(N).
6. Além disso, a co-produção(N) com o canal(N) francês TF1(N) para a realização(N) de mais 30 episódios(N) continua(V) sendo(V) negociada(V).
7. Sob o comando(N) de Ronaldo(N) Rosas(N), o programa(N) mostrará(V) reportagens(N) especiais de Sônia(N) Pompeu(N).
8. A direção(N) do novo semanal(N) será(V) assinada(V) por Ewaldo(N) Ruy(N).
9. Os jogadores(N) se dividem(V) pelos dez quartos(N) do alojamento(N), equipados(V) com frigobar(N), ar(N) condicionado, televisão(N) e telefone(N).
10. Além de Maurício(N), Carlão(N) e Paulão(N), a seleção(N) deve(V) contar(V) hoje com Giovane(N).
11. O atacante(N), que deveria(V) ter(V) se apresentado(V) anteontem à noite(N), pediu(V) mais um dia(N) de folga(N) ao treinador(N).
12. Na volta(N) de uma viagem(N) ao exterior(N), vale(V) a pena(N) trazer(V) uma impressora(N) matricial.
13. Free shops dos aeroportos(N) internacionais também vendem(V) o equipamento(N).

14. O modelo(N) LX 810, da Epson(N), é(V) vendido(V) em Miami(N) por US$ 178.
15. O preço(N) de lista(N) nas revendas(N) brasileiras é(V) de US$ 422.
16. Esse equilíbrio(N) era(V) tido(V) como pré-condição(N) para o sucesso(N) do plano(N) econômico.
17. **Parte(V)***[N]* dos recursos(N) para a formação(N) do FSE(N) foi(V) deslocado(V) do orçamento(N) da saúde(N) e educação(N).
18. Na época(N), o então ministro(N) da Fazenda(N), Fernando(N) Henrique(N) Cardoso(N), fez(V) um pronunciamento(N) em cadeia(N) nacional para anunciar(V) a intenção(N) do governo(N) de destinar(V) o FSE(N) a investimentos(N) sociais.
19. O assessor(N) de imprensa(N) do Ministério(N) da Fazenda(N), Sérgio(N) Danese(N), disse(V) ontem que o ministro(N) da Fazenda(N), Rubens(N) Ricúpero(N) não iria(V) comentar(V) o assunto(N) porque não tinha(V) informações(N) suficientes.
20. O projeto(N) original do governo(N) destinava(V) ao TSE(N) R$ 334, 9 milhões.
21. Como não houve(V) acordo(N) entre governo(N) e tribunal(N) quanto ao volume(N) de recursos(N), a dotação(N) foi(V) incluída(V) na reserva(N) de contingência(N) sem especificação(N) de despesa(N).
22. Posteriormente, diante da ameaça(N) do tribunal(N) de entrar(V) com uma ação(N) judicial, o governo(N) mandou(V) ao Congresso(N) uma alteração(N) ao projeto(N), aumentando(V) para R$ 452, 7 milhões a dotação(N) do TSE(N).
23. Essas medidas(N) reduziram(V) a disponibilidade(N) de dinheiro(N) no sistema(N) bancário e os grandes bancos(N) passaram(V) a não fornecer(V) recursos(N) para as pequenas instituições(N).
24. Aqui só joga(V) quem está(V) bem.

25. Ninguém força(V) sua escalação(N) porque há(V) quem escale(V) o time(N) no São(N) Paulo(N).
26. Ele só não jogava(V) porque não estava(V) bem.
27. Apenas dois árbitros(N) resolveram(V) contar(V) todos os podres(N), enquanto a federação(N) tem(V) mais de 70.
28. O futebol(N) precisa(V) seguir(V) o exemplo(N) da CPI(N) do orçamento(N) e apresentar(V) todos os podres(N).
29. Se eu dirigisse(V) uma federação(N), apresentaria(V) balanços(N) mensais e liberaria(V) minhas contas(N) bancárias.
30. A Fifa(N) e a CBF(N) deveriam(V) entrar(V) de sola(N) nesse caso(N) e em todas as outras federações(N).
31. A maioria(N) das empresas(N) que produzem(V) leite(N) das marcas(N) interditadas(V)*[ADJ]* não tinha(V) sido(V) comunicada(V) ontem sobre a liberação(N) do produto(N).
32. O presidente(N) da Cooper(N), Benedito(N) Vieira(N) Pereira(N), 49, afirmou(V) que pretendia(V) distribuir(V) leite(N) C(N) nos postos(N) de venda(N) hoje.
33. O Applause(N), um sedã(N) quatro portas(N), com motor 1.6, é(V) o carro(N) mais caro da Daihatsu(N).
34. O top(N) de linha(N) custa(V) US$ 30 mil.
35. A mudança(N) do local(N) de jogo(N) que deve(V) acontecer(V) também na partida(N) contra o Corinthians(N), no próximo dia(N) 17 foi(V) determinada(V) pela CBF(N), que não viu(V) garantias(N) de segurança(N) no estádio(N) santista.
36. Souza(N) também negou(V) aos réus(N) o direito(N) de apelarem(V) da setença(N) em liberdade(N).

37. Os três estão(V) presos(N)*[ADJ]* desde 30 de julho(N) de 93.
38. Começou(V) bem antes do que se previa(V) a batalha(N) pela futura sucessão(N) na Fifa(N) apenas seis meses(N) depois do super-acordo(N) que, nas vésperas(N) da Copa(N), reconduziu(V) o brasileiro João(N) Havelange(N) ao sexto mandato(N) consecutivo.
39. Pelo acordo(N), os três continentes(N) mais obstinados em cortar(V) o reinado(N) de Havelange(N), a África(N), a Ásia(N) e a Europa(N), aceitaram(V) cancelar(V) os seus movimentos(N) de oposição(N) em troca(N), basicamente, de dois compromissos(N).
40. Havelange(N) aceitaria(V) engolir(V) o italiano Antonio(N) Matarrese(N) como o seu vice-executivo(N) e, além disso, esqueceria(V) os seus modos(N) autoritários, coordenando(V) a entidade(N) de maneira(N) colegiada(N)[*ADJ*].
41. O Ambulim(N) foi(V) um dos centros(N) que contribuíram(V) para um estudo(N) apresentado(V) na 5ª Conferência(N) Internacional(N) sobre Transtornos(N) Alimentares(N), de 29 de abril(N) a 1º de maio(N) em Nova(N) York(N).
42. **Dados(V)***[N]* sobre abuso(N) sexual em bulímicas(N) no Brasil(N), Áustria(N) e Estados(N) Unidos(N) foram(V) centralizados(V) por Harrison(N) Pope(N), da Escola(N) de Medicina(N) de Harvard(N).
43. Essa divisão(N) gera(V) algumas distorções(N) terríveis.
44. Afora historicismo(N), isso é(V) menosprezar(V) um fator(N) interno à arte(N) brasileira, que independe(V) de contexto(N) internacional.
45. Volpi(N) foi(V) dos mais influentes pintores(N) do país(N) para além(N)*[ADV]* da questão(N) da autonomia(N).
46. O panorama(N) sofre(V) prejuízos(N) demais em favor(N) da tese(N).

47. Abstratos(N) entre medianos(N) e medíocres(N), como Fukushima(N), Pérsio(N), Raimo(N) e Douchez(N), têm(V) o mesmo(N)*[ADJ]* ou maior destaque(N) que Volpi(N) e nada que se possa(V) chamar(V) de autonomia(N) para oferecer(V) como lenitivo(N).
48. Talvez isto seja(V) muito barulho(N) por nada.
49. Original(N): Cícero(N).
50. Disse(V) que não conseguia(V) vislumbrar(V) artifícios(N) fraudulentos ou prática(N) de peculato(N) no protocolo(N) assinado(V) por Quércia(N).
51. Afirmou(V) que o conjunto(N) de fatos(N), em princípio(N), aponta(V) o envolvimento(N) de Quércia(N).
52. Recebeu(V) a denúncia(N).
53. Segundo o médico(N), o caso(N) não preocupa(V).
54. Romário(N) não se exercitou(V) nas cobranças(N) de falta(N) e pênalti(N).
55. Toda a comissão(N) técnica sabe(V) que Romário(N) é(V) de treinar(V) pouco, geralmente se poupando(V) entre dois jogos(N) difíceis.
56. Nem o PSB(N) nem a coligação(N) têm(V) competência(N) legal para trocar(V) o vice(N) da chapa(N) sem a concordância(N) de Bisol(N), que teve(V) o nome(N) aprovado(V) em convenção(N).
57. Outra maneira(N) de um partido(N) forçar(V) a substituição(N) seria(V) expulsar(V) o candidato(N), com base(N) em seu estatuto(N).
58. Neste caso(N), o registro(N) da candidatura(N) seria(V) cancelado(V) pela Justiça(N).
59. Apesar de limitar(V) a venda(N) de quatro ingressos(N) por pessoa(N), a Mesbla(N) não evitava(V) ontem que uma mesma pessoa(N) comprasse(V) mais de uma vez(N).
60. A queda(N) nas vendas(N) teve(V) reflexos(N) nas negociações(N) entre as confecções(N) e as lojas(N).

61. A seca(N) que atingiu(V) as áreas(N) produtoras de grãos(N) não deve(V) causar(V) grandes estragos(N) na safra(N) 1994/95.
62. A primeira previsão(N) do IBGE(N) (Fundação(N) Instituto(N) Brasileiro(N) de Geografia(N) e Estatística(N)) indica(V) queda(N) de 0,62% na área(N) plantada(V)*[ADJ]* nesta safra(N) em relação(N) a anterior.
63. Segundo ele, a exposição(N) ao material(N) durante a gravidez(N) ou nos dois primeiros anos(N) de vida(N) não representa(V) perigo(N).
64. Foi(V) utilizada(V) técnica(N) mista, incluindo(V) desenho(N), pastel(N), ceragrafia(N) e fotografismo(N).
65. Monica(N) Filgueiras(N) Galeria(N) de Arte(N) (al. Min.(N) Rocha(N) Azevedo(N), 927, tel. 282-5292).
66. De seg(N) a sex(N) das 11h às 19h, sáb(N) das 11h às 14h.
67. Até 30 de março(N).
68. Preço(N) das obras(N) : de US$ 2.000 a US$ 4.000.
69. Não se trata(V) simplesmente da questão(N) de a escravidão(N) certamente ter(V) tido(V) efeitos(N) duradouros sobre a cultura(N) negra, nem mesmo dela ter(V) exercido(V) um amplo efeito(N) negativo sobre a auto-confiança(N) e auto-estima(N) dos negros(N), mas mais especificamente de a experiência(N) da escravatura(N) ter(V) desvirtuado(V) e tolhido(V) a evolução(N) do algoritmo(N) etnocêntrico que os negros(N) americanos teriam(V) desenvolvido(V) no **decorrer(V)***[N]* normal(N)*[ADJ]* dos acontecimentos(N).
70. Os brancos(N) fizeram(V) tudo em seu poder(N) para invalidar(V) ou menosprezar(V) cada sinal(N) de talento(N), virtude(N) ou superioridade(N) entre os negros(N).

71. Eles tiveram(V) que fazer(V) isso se os escravos(N) fossem(V) superiores(N)*[ADJ]* em qualidades(N) que os próprios brancos(N) valorizavam(V), onde estaria(V) a justificativa(N) moral para mantê-los(V) escravizados(V)*[ADJ]*?
72. E, assim, tudo o que os afro-americanos(N) faziam(V) bem teve(V) de ser(V) colocado(V) em termos(N) que menosprezassem(V) a qualidade(N) em questão(N).
73. Mesmo a simples tentativa(N) de se documentar(V) esse ponto(N) deixa(V) uma pessoa(N) exposta*[V]* a acusações(N) de condescendência(N) e, assim, os brancos(N) de fato(N) conseguiram(V) cooptar(V) os julgamentos(N) de valor(N).
74. É(V) ainda mais óbvio que é(V) impossível(N)*[ADJ]* falar(V) abertamente sobre o superioridade(N) de muitos atletas(N) negros sem ser(V) sujeito(N)*[ADJ]* a acusações(N) de que se estar(V) sendo(V) anti-negro(N)*[ADJ]* de uma maneira(N) enviesada.
75. Pela segunda vez(N) desde quando começou(V) a coordenar(V) as ações(N) no Rio(N), há(V) duas semanas(N), o Exército(N) mudou(V) o nome(N) das operações(N).
76. Agora, os oficiais(N) envolvidos(V)*[ADJ]* se referem(V) sempre ao comando(N) das ações(N) como Centro(N) de Coordenação(N) de Operações(N) de Combate(N) ao Crime(N) Organizado(N) (Ccocco(N)).
77. Cinco linhas(N) paralelas, de mais de 400 km cada, foram(V) **descobertas(N)***[V]* por cientistas(N) australianos no sul(N) do país(N).
78. Elas estão(V) separadas(V) por espaços(N) de 80 km a 100 km.
79. As linhas(N), invisíveis(N)*[ADJ]* da superfície(N), foram(V) detectadas(V) através de dados(N) de satélites(N).
80. Pesquisadores(N) acham(V) que as linhas(N) podem(V) ser(V) falhas(N) geológicas.

81. O derramamento(N) de óleo(N) do **petroleiro***[N]* Exxon(N) Váldez(N), em março(N) de 1989, causou(V) estragos(N) no valor(N) de US$ 286,8 milhões, segundo um júri(N) em Anchorage(N) (Alaska(N)).
82. A ação(N) está(V) sendo(V) movida(V) por pescadores(N), lojistas(N), propietários(N) de terra(N) e nativos(N).
83. O valor(N) é(V) mais que o dobro(N) do estimado(V) pela Exxon(N), mas menor(N)*[ADJ]* que o original(N) pedido, de US$ 895 milhões.
84. Bombeiros(N) e pessoal(N) de resgate(N) foram(V) colocados(V) em **alerta***[N]* máximo(N)*[ADJ]* em Argel(N) no final(N) da tarde(N) de ontem(N)*[ADV]*, segundo a rádio(N) estatal.
85. Veículos(N) de resgate(N) estavam(V) a apenas 500 metros(N) do Airbus(N) 300.
86. Optar(V) entre um aparelho(N) conjugado e outro simples é(V) outro ponto(N) que merece(V) atenção(N).
87. Para fazer(V) uma boa compra(N), a técnica(N) do Procon(N) recomenda(V) ao consumidor(N) que verifique(V) se já há(V) dentro de casa(N) os tradicionais equipamentos(N) que desempenham(V) as mesmas funções(N) do conjugado(N), só que separadamente.
88. **Caso(V)***[N]* a opção(N) seja(V) pelo aparelho(N) multiuso, o comprador(N) deve(V) checar(V) se o produto(N) tem(V) assistência(N) técnica, diz(V) ela.
89. Como(V)*[CONJ]* a idéia(N) de enxugar(V) a Constituição(N) enfrenta(V) resistência(N) inclusive nos partidos(N) que apóiam(V) o governo(N), a equipe(N) de FHC(N) resolveu(V) fazer(V) as reformas(N) por partes(N).
90. Primeiro aprova-se(V) o texto(N) enxuto e depois negocia-se(V) a aprovação(N), sem prazo(N) definido, das leis(N) complementares e ordinárias(N)*[ADJ]*.

91. O Pentágono(N) usa(V) a Internet(N), que conecta(V) computadores(N) a sistemas(N) telefônicos, para que seus funcionários(N) troquem(V) informações(N).
92. Os arquivos(N) são(V) protegidos(V) por senhas(N) e códigos(N), que acabaram(V) mostrando-se(V) vulneráveis.
93. De Mario(N) Bernardini(N), da Fiesp(N), sobre a seleção(N) e o real(N): se reúnem(V) entre sexta-feira(N) e domingo(N) em Araxá(N) (MG(N)).
94. É(V) o 5º Encontro(N) Nacional(N) de Histórias(N) em Quadrinhos(N).
95. Haverá(V) oficinas(N) e um concurso(N).
96. Informações(N) pelo telefone(N) (034) 661-2458.
97. Para desenhistas(N) é(V) o tema(N) da oficina(N) com David(N) Campitti(N) (veja(V) trabalho(N) dele ao lado(N)), roteirista(N) de histórias(N) do Super-Homem(N) e criador(N) de vários personagens(N).
98. Outros profissionais(N) brasileiros, que atuam(V) nos EUA(N), também participam(V).
99. Informações(N) pelo (011) 263-4700.
100. Ontem pela manhã(N), os atletas(N) que não atuaram(V) contra o Vitória(N), participaram(V) de um coletivo(N) com a equipe(N) principal.
101. Os que jogaram(V), fariam(V) treino(N) de recuperação(N) à tarde(N) no Projeto(N) Acqua(N).
102. Murici(N) reclamou(V) da postura(N) do time(N) do Vitória(N).
103. Pereira(N) levou(V) um pisão(N) na cabeça(N), mesmo estando(V) caído(V) no chão(N).
104. Murici(N) espera(V) que o Inter(N) não tenha(V) a mesma atitude(N).
105. A primeira missão(N) lunar dos EUA(N) em 21 anos(N) teve(V) início(N) ontem(N)*[ADV]*, quando um foguete(N) Titã(N) 2 colocou(V) no espaço(N) a astronave(N) não tripulada(V)*[ADJ]* Clementine(N) 1, que vai(V) passar(V) dois

meses(N) em duas órbitas(N) da Lua(N) para realizar(V) completo mapeamento(N) mineralógico e topográfico do satélite(N) da Terra(N).

106. O nome(N) oficial do projeto(N) é(V) Depse(N) 1.

107. Ele é(V) patrocinado(V) pela Organização(N) de Defesa(N) de Mísseis(N) Balísticos(N) e pela Nasa(N), numa das primeiras operações(N) espaciais com fins(N) civis e militares(N)*[ADJ]*.

108. O lançamento(N) ocorreu(V) ontem na base(N) aérea de Vanderberg(N), na Califórnia(N), costa(N) oeste do país(N), às 8h30 locais (14h30 em Brasília(N)).

109. O custo(N) total do projeto(N) é(V) de U$ 55 milhões.

110. O atual perfil(N) dos poupadores(N) ajudará(V) a manter(V) o dinheiro(N) aplicado.

111. Segundo o BC(N), mais da metade(N) dos recursos(N) da caderneta(N) são(V) de poupadores(N) de médio e grande(N)*[ADJ]* **porte(V)***[N]* em tese(N), menos sujeitos a achar(V) que o dinheiro(N) perdeu(V) rendimento(N) com a queda(N) da inflação(N).

112. A previsão(N) de saques(N) reduzidos(V)*[ADJ]* sobre a poupança(N) é(V) compartilhada(V) por especialistas(N).

113. Diniz(N) começou(V) sua carreira(N) automobilística em 1989, no Brasileiro(N) de Fórmula(N) Ford(N), campeonato(N) em que obteve(V) a sexta posição(N) na classificação(N) final.

114. A Polícia(N) Civil(N) de Ourinhos(N) (371 km a oeste(N) de São(N) Paulo(N)) prendeu(V) ontem à tarde(N) o ex-líder(N) religioso Jonas(N) Rúbio(N), 45, acusado(V) de matar(V) na quarta-feira(N) a **estudante***[N]* Claudirene(N) Contijo(N), 13, com um tiro(N) de espingarda(N).

115. O **delegado***[N]* Celso(N) Antonio(N) Borlina(N), 38, disse(V) que Rúbio(N) confessou(V) o crime(N).

116. Rúbio(N) era(V) acusado(V) por Claudirene(N) de tê-la(V) estuprado(V) no ano(N) passado, época(N) em que era(V) o líder(N) da Assembléia(N) de Deus(N) na usina(N) São(N) Luiz(N), onde a estudante(N) morava(V).