Análise de Expressões Referenciais em Corpus Anotado da Língua Portuguesa

Sandra Collovini de Abreu

São Leopoldo
2005

UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA – PIPCA

Sandra Collovini de Abreu

Análise de Expressões Referenciais em Corpus Anotado da Língua Portuguesa

Dissertação de Mestrado submetida a avaliação como requisito parcial para obtenção do grau de Mestre em Computação Aplicada

Orientadora:
Profª. Drª. Renata Vieira

São Leopoldo
2005

Dedico este trabalho em memória à minha mãe que infelizmente não pode acompanhá-lo, mas pode contribuir muito com o seu exemplo de determinação, amor, fé e luta pela realização de seus sonhos.

Agradecimentos

À minha orientadora, Profª. Drª. Renata Vieira, pela dedicação, comprometimento, confiança e amizade durante todo o Mestrado.

À meu grande amor, James, por todo o companheirismo, incentivo, compreensão e amor. À minha sogra Clenir e sogro João por todo o apoio.

À toda a minha família, pelo amor, compreensão, incentivo, em especial a minha irmã Ângela, que me acompanhou durante toda esta etapa.

Aos meus colegas do laboratório de Engenharia da Linguagem, Cássia, Cassiana, Cláudia, César, Douglas, Fábio, Leonardo, Rodrigo, Sandro, Vinícius, pela amizade e auxílio no transcorrer deste estudo. Em especial aos colegas César, por todo o apoio e pela realização da anotação de correferência no corpus Público, e ao Rodrigo pelo auxílio na elaboração das folhas de estilo.

Aos colegas de Portugal, Ana Margarida Aires pela contribuição no processo de comparação da classificação manual e automática e ao Prof. Dr. Paulo Quaresma pelo acompanhamento e cooperação.

Aos colegas do Mestrado, por todo o companheirismo. Em especial aos colegas Adriana, Daniela, Gabriela, Letícia, Isa Mara e Rafael Torchelsen, por todos os momentos de ajuda e incentivo.

Ao corpo docente do PIPCA pela disposição; aos administradores de rede, Fábio e Antônio pela prestatividade. À secretaria Rejane, por todo o auxilio e amizade.

À CAPES pelo auxílio financeiro durante o transcorrer do Mestrado e pelo financiamento do projeto DIRPI de cooperação internacional com Portugal, pela oportunidade de visitar à Universidade de Évora e aprofundar os estudos relacionados a este trabalho.

Resumo

A análise de expressões referenciais é fundamental na interpretação do discurso. A identificação de expressões correferentes é importante em diversas aplicações de Processamento da Linguagem Natural. Expressões referenciais podem ser usadas para introduzir entidades em um discurso ou podem fazer referência a entidades já mencionadas, podendo fazer uso de redução lexical, como em: *"O Eurocenter oferece cursos de Japonês na bela cidade de Kanazawa. Os cursos têm quatro semanas de duração"*. Onde "*cursos de Japonês"* introduz uma nova entidade e "o*s cursos"* retomam essa entidade. A resolução de correferência é o processo de identificar as expressões que se referem à mesma entidade no discurso. As expressões referenciais são analisadas e a existência de um antecedente textual é verificada. Aquelas que introduzem novos elementos, chamamos novas no discurso.

Esta dissertação apresenta um estudo das características de um tipo específico de expressões referenciais (descrições definidas) com o objetivo de identificar automaticamente expressões novas no discurso em textos da Língua Portuguesa. Este estudo é importante, pois o número de expressões sem antecedentes textuais no discurso tanto na Língua Inglesa como na Língua Portuguesa é expressivo.

O estudo das características baseou-se na literatura e em um estudo de corpus. A partir destas características foi construída uma base de dados para o aprendizado automático de árvores de decisão. Os melhores resultados da classificação das descrições definidas foram implementados no ambiente ART. Uma análise dos atributos foi desenvolvida para calcular o potencial de distinção de cada um, destacando-se o atributo "*tamanho*" (número de palavras do sintagma nominal) por ser um atributo original e significativo nos experimentos e o

atributo "*sem antecedente*" (núcleo da descrição definida é uma palavra que não ocorre anteriormente no texto) por ter um impacto positivo nos resultados.

As árvores de decisão geradas foram avaliadas em um novo corpus, composto por textos extraídos do jornal português Público. Obtivemos 77% de F-measure para a identificação de expressões novas no discurso.

Palavras chave: expressões referenciais, classificação automática de expressões referenciais, resolução de correferência, resolução de anáforas, aprendizado de máquina.

Abstract

The analysis of referring expressions is fundamental to discourse interpretation. The identification of correfering expressions is an important step in many Natural Language Processing applications. Referring expressions may introduce new entities in a discourse or the can refer back to already mentioned entities; they can do it on the basis of simplified expressions, as in "*Eurocenter offers Japanese courses in Kanazawa. The courses are four week long*." In this example, the expression "*Japenese courses"* introduces a new discourse entity, whereas "*the courses*" refer back to this already mentioned entity.

Correference resolution is the process of identification expressions that refer to the same entity. Referring expressions are analysed and the existence of a textual antecedent is verifyed. Those which introduce new discourse entities are considered discourse new in the discourse.

This dissertation presents a study of a specific type of referring expression (definite description) with the goal of identifying discourse new expressions in Portuguese texts. This is an important issue since the number of such expressions without textual antecedents were found to be significant both in English and Portuguese corpora.

The study of definite description features was based both on the literature and on corpus studies. A data base for learning decisions trees was constructed. The generated trees through the learning process were implemented and evaluated in the ART framework.

The features were analysed individually and we found that a new atribute based on the size of the noun phrase presented interesting results. Another relevant atribute with good impact in the results is based on the lack of the head noun in the previous text.

The automatically generated trees were evaluated in a new corpus, composed of European Portuguese texts from the Portuguese newspaper Publico. We had an F-measure of 77% for the identification of discourse new.

Lista de Figuras

FIGURA 2.1 ESTRUTURA DA ÁRVORE DE DECISÃO 27
FIGURA 3.1 ARQUIVO DE *WORDS*. 40
FIGURA 3.2 ARQUIVO DE *POS*. 40
FIGURA 3.3 ARQUIVO DE *CHUNKS*. 40
FIGURA 3.4 ARQUIVO DA ESTRUTURA. 42
FIGURA 3.5 ARQUIVO DE MARCAÇÕES. 42
FIGURA 3.6 ARQUITETURA DA FERRAMENTA ART (GOULART, GASPERIN; VIEIRA, 2004) 44
FIGURA 3.7 ARQUIVO DE ANAFÓRICAS 45
FIGURA 3.8 ARQUIVO DE CANDIDATOS A ANTECEDENTES 45
FIGURA 3.9 ARQUIVO DE MARCAÇÕES 45
FIGURA 3.10 ARQUIVO NO FORMATO ARFF 46
FIGURA 3.11 ABRANGÊNCIA 48
FIGURA 3.12 PRECISÃO 48
FIGURA 3.13 F-MEASURE 49
FIGURA 3.14 TRECHO DO ARQUIVO DE *CHUNKS* 55
FIGURA 3.15 TRECHO DO ARQUIVO DE *CHUNKS* 55
FIGURA 3.16 TRECHO DO ARQUIVO DE *CHUNKS* 56
FIGURA 3.17 METODOLOGIA PROPOSTA 57
FIGURA 3.18 ARQUIVO DE MARCAÇÕES 59
FIGURA 3.19 ARQUIVO DE *CHUNKS* 59
FIGURA 3.20 TRECHO DE UM TEXTO 59
FIGURA 3.21 LISTA DE DESCRIÇÕES DEFINIDAS 59
FIGURA 3.22 TRECHO DO ARQUIVO DE ENTRADA NO FORMATO ARFF 62
FIGURA 4.1 ÁRVORE DE DECISÃO COM ATRIBUTOS G1 75
FIGURA 4.2 ÁRVORE DE DECISÃO COM ATRIBUTOS G1 SEM TAM 76

Lista de Tabelas

TABELA 2.1 CONFIGURAÇÕES DAS DESCRIÇÕES DEFINIDAS. 25
TABELA 2.2 ALGUMAS CARACTERÍSTICAS DE TRABALHOS RELACIONADOS. 36
TABELA 3.1 INFORMAÇÕES SOBRE O CORPUS 1. 37
TABELA 3.2 INFORMAÇÕES SOBRE O CORPUS 2. 38
TABELA 4.1 NÚMERO DE EXEMPLOS DE CADA CLASSE. 67
TABELA 4.2 RESULTADOS DO *BASELINE*. 67
TABELA 4.3 EXPERIMENTO 1 COM ATRIBUTOS G1. 68
TABELA 4.4 RESULTADOS DO EXPERIMENTO 1 COM ATRIBUTOS G12. 68
TABELA 4.5 RESULTADOS DO EXPERIMENTO 1 COM G123 DE ATRIBUTOS. 69
TABELA 4.6 PERCENTUAL DE ACERTOS DO EXPERIMENTO 1 69
TABELA 4.7 COMPARAÇÃO ENTRE O *BASELINE* E O EXPERIMENTO 1 70
TABELA 4.8 NÚMERO DE EXEMPLOS DE CADA CLASSE. 70
TABELA 4.9 RESULTADOS DO *BASELINE*. 70
TABELA 4.10 RESULTADOS DO EXPERIMENTO 2 COM ATRIBUTOS G1. 71
TABELA 4.11 RESULTADOS DO EXPERIMENTO 2 COM ATRIBUTOS G12. 71
TABELA 4.12 RESULTADOS DO EXPERIMENTO 2 COM ATRIBUTOS G123. 72
TABELA 4.13 PERCENTUAL DE ACERTOS DO EXPERIMENTO 2. 72
TABELA 4.14 COMPARAÇÃO ENTRE O *BASELINE* E EXPERIMENTO 2. 73
TABELA 4.15 POTENCIAL DE DISTINÇÃO DOS ATRIBUTOS G1 (CF = 0,35) 74
TABELA 4.16 POTENCIAL DE DISTINÇÃO DOS ATRIBUTOS G12 (CF = 0,35). 74
TABELA 4.17 POTENCIAL DE DISTINÇÃO DOS ATRIBUTOS G123 (CF = 0,10) 74
TABELA 4.18 COMPARAÇÃO DO ATRIBUTO TAM 75
TABELA 4.19 POTENCIAL DE DISTINÇÃO DOS ATRIBUTOS G1 (CF = 0,35) 76
TABELA 4.20 POTENCIAL DE DISTINÇÃO DOS ATRIBUTOS G12 (CF = 0,35). 77
TABELA 4.21 POTENCIAL DE DISTINÇÃO DOS ATRIBUTOS G123 (CF = 0,35). 77
TABELA 4.22 *PIPES* E FILTROS DO EXPERIMENTO 1 78
TABELA 4.23 RESULTADO COMPARATIVO DO EXPERIMENTO 1 78
TABELA 4.24 *PIPES* E FILTROS DO EXPERIMENTO 2. 79
TABELA 4.25 RESULTADO COMPARATIVO DO EXPERIMENTO 2 80
TABELA 4.26 NÚMERO DE EXEMPLOS POR CLASSE DOS EXPERIMENTOS 80
TABELA 4.27 RESULTADOS DOS *BASELINES* DOS EXPERIMENTOS. 81
TABELA 4.28 RESULTADO COMPARATIVO DO EXPERIMENTO 1 81
TABELA 4.29 RESULTADO COMPARATIVO DO EXPERIMENTO 2 82
TABELA 5.1 RESULTADOS DOS TRABALHOS RELACIONADOS. 86

Lista de Abreviaturas

AM	Aprendizado de Máquina
ART	*Anaphor Resolution Tool*
IA	Inteligência Artificial
MMAX	*Multi-Modal Annotation in XML*
NILC	Núcleo Interinstitucional de Lingüística Computacional
PLN	Processamento de Linguagem Natural
POS	*Part of Speech*
RI	Recuperação de Informação
SCT	Sistema de Categorização de Textos
Weka	*Waikato Environment for Knowledge Analysis*
XML	*eXtensible Markup Language*
XSL	*eXtensible Stylesheet Language*

Sumário

1. Introdução 14

1.1 Objetivos 15
1.1.1 Objetivo Geral 15
1.1.2 Objetivos Específicos 15
1.2 Organização do Texto 16

2. Revisão bibliográfica 18

2.1 Expressões Referenciais: Sintagmas Nominais e Descrições Definidas 18
2.2 Correferência e Anáfora 19
2.3 Classificação das descrições definidas 21
2.4 Estudo das características das descrições definidas 22
2.4.1 Partes fundamentais das descrições definidas 23
2.4.2 Os complementos das descrições definidas 23
2.5 Árvores de Decisão 25
2.6 Trabalhos Relacionados 29
2.7 Considerações Finais 35

3. Materiais e Métodos 37

3.1 Descrição do Corpus 37
3.2 PALAVRAS e Xtractor 39
3.3 MMAX 41
3.4 ART 43
3.5 Weka 45
3.6 Avaliação 47
3.7 Grupos de Características das descrições definidas 49
3.8 Identificação Automática das Características no Corpus 54
3.9 Processo de Aprendizado 57
3.9.1 Coleta da Base 57
3.9.2 Pré-processamento 57
3.9.3 Classificação 61
3.10 Implementação no Ambiente ART 62
3.11 Considerações Finais 63

4. Resultados 66

4.1 Geração de Árvores de Decisão 66
4.1.1 Experimento 1 – novas no discurso (Weka) 67
2.1.2 Experimento 2 – não correferentes (Weka) 70
4.1.3 Potencial de Distinção das Características 73
4.2 Avaliação das Árvores no Ambiente ART 77
4.2.1 Avaliação das Características em um novo corpus 80
4.3 Considerações finais 82

5. Conclusões e Trabalhos Futuros 84

Referências **87**
APÊNDICE A – Folhas de Estilo XSL **92**
APÊNDICE B – Árvores de Decisão **101**

1. Introdução

Em sistemas de Processamento de Linguagem Natural (PLN), a análise de expressões referenciais é um componente fundamental na interpretação do sentido do texto. A resolução de correferência constitui do ponto de vista do processamento computacional, um problema difícil, devido à complexidade do fenômeno lingüístico e por isso motiva inúmeras pesquisas.

O processo de resolução de correferência busca identificar expressões lingüísticas que se referem à mesma entidade (correferentes). Para exemplificar, segue o trecho de um texto, onde as expressões correferentes estão destacadas.

"O advogado de Castor de Andrade, Nélio Machado, afirmou que vai aguardar a evolução dos fatos para se pronunciar. O advogado disse desconhecer a existência de documentos que demonstrariam o pagamento de propinas a autoridades policiais".

A identificação de expressões correferentes é importante em diversas aplicações de PLN, como, por exemplo, em sumarização automática, extração de informação, recuperação de informação, tradução automática, classificação de textos, entre outros.

Este estudo trata de expressões referenciais em textos da Língua Portuguesa, especificamente de descrições definidas. Chamamos descrições definidas os sintagmas nominais iniciados por artigo definido (o, a, os, as). As descrições definidas são estudadas extensivamente pela lingüística, filosofia, psicologia e lingüística computacional (VIEIRA, 1998). Além disso, trabalha-se com descrições definidas pelo fato de ocorrerem em grande quantidade nos textos. Também, existem vários trabalhos sobre resolução de anáforas pronominais, contudo existem poucos trabalhos que abordam a resolução anafórica das descrições definidas para a Língua Portuguesa.

Estudos anteriores (VIEIRA; GASPERIN; GOULART, 2003) mostram que as descrições definidas em textos jornalísticos possuem um antecedente textual em apenas 50% dos casos. Esse fato aliado à complexidade da tarefa de encontrar um antecedente para resolução de correferência estimula a comunidade a propor como parte do processo de resolução, a identificação de descrições definidas novas no discurso (MCCARTHY, LEHNERT, 1995; BEAN, RILOFF, 1999; CARDIE; WAGSTAFF,1999; VIEIRA, POESIO, 2000; (SOON; NG; LIM, 2001; MULLER; RAPP; STRUBE, 2002; NG, CARDIE, 2002a, 2002b; STRUBE; RAPP; MULLER, 20002; URYUPINA, 2003; POESIO et al., 2005). Todos esses trabalhos são estudos da Língua Inglesa. Nessa dissertação realizamos um estudo detalhado das expressões referenciais definidas da Língua Portuguesa e das suas características para a identificação daquelas que não possuem um antecedente textual.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho tem como objetivo geral o estudo de características das descrições definidas em textos da Língua Portuguesa para a sua classificação como expressões novas no discurso. Para isso, uma análise das ocorrências de descrições definidas em textos da Língua Portuguesa é realizada. A abordagem de Aprendizado de Máquina Supervisionado (árvores de decisão) para a avaliação das características relevantes é utilizada. Por fim, uma análise das árvores de decisão geradas é realizada para posterior implementação no ambiente ART.

1.1.2 Objetivos Específicos

- Estudo das descrições definidas nos textos da Língua Portuguesa, com base em corpus anotado com relações de correferência;

- Levantamento de características para a classificação das descrições definidas como expressões novas no discurso (não possuem um antecedente textual) ou outra (possuem um antecedente textual) com base no estudo de corpus e na literatura disponível para a Língua Inglesa (MCCARTHY, LEHNERT, 1995; BEAN, RILOFF, 1999; CARDIE, WAGSTAFF,1999; VIEIRA, POESIO, 2000; SOON; NG; LIM, 2001; MULLER; RAPP; STRUBE, 2002; NG, CARDIE, 2002a; NG, CARDIE, 2002b; STRUBE et al, 20002; URYUPINA, 2003; POESIO et al., 2005);
- Construção da base de dados para o aprendizado automático, com base nas características identificadas que servirão de atributos para as árvores de decisão;
- Geração e análise de árvores de decisão para a investigação dos atributos relevantes na classificação binária das descrições definidas como nova no discurso e outra;
- Implementação das árvores de decisão no ambiente ART (GOULART; GASPERIN; VIEIRA, 2004);
- Avaliação final das árvores de decisão implementadas no ambiente ART com um novo corpus.

1.2 Organização do Texto

Esta dissertação está organizada da seguinte forma. No capítulo 2, é dada uma introdução aos conceitos importantes relacionados a esse trabalho. Um estudo inicial das descrições definidas é apresentado. Após, uma visão geral de Árvores de Decisão é mostrada. Por fim, os trabalhos relacionados.

O capítulo 3 apresenta uma visão geral dos materiais e métodos da pesquisa: corpus e ferramentas, além das medidas de avaliação que serão utilizadas nos experimentos. Uma análise detalhada das características morfossintáticas das descrições definidas do corpus anotado é mostrada. Após, o processo para a identificação automática das características para

geração da base de dados do aprendizado, utilizando a linguagem XSL é abordado. Por fim, o processo de aprendizado com árvores de decisão e as etapas para a implementação das árvores de decisão no ambiente ART são apresentados.

No capítulo 4, são apresentados os resultados da geração das árvores e posterior implementação das mesmas no Ambiente ART. A avaliação dos experimentos no ambiente ART com um novo corpus é mostrada.

Por fim, no capítulo 5 são apresentadas as conclusões e uma discussão sobre os trabalhos futuros.

2. Revisão bibliográfica

O objetivo deste capítulo é abordar conceitos importantes das áreas de lingüística e computação relacionados à dissertação.

Com base na literatura, é apresentado o conceito de descrições definidas, as estruturas de interesse deste trabalho, assim como as configurações possíveis que estas podem adotar. Além disso, é exposta a classificação das descrições definidas empregada neste trabalho. Um estudo das características das descrições definidas usadas na classificação é apresentado. Este capítulo também aborda Árvores de Decisão. Os principais trabalhos relacionados que contribuíram para este estudo são referenciados.

2.1 Expressões Referenciais: Sintagmas Nominais e Descrições Definidas

Um sintagma é uma palavra ou um conjunto de palavras, que constituem uma unidade significativa dentro da sentença (BONINI, 2002, KOCH, 2003; MACAMBIRA, 1990).

Os sintagmas desempenham diferentes funções na sentença e combinam-se em torno de um núcleo. É o núcleo que denomina o sintagma. O sintagma pode ser nominal (núcleo nome ou pronome), verbal (núcleo verbo), preposicional (núcleo preposição), adjetival (núcleo adjetivo) e adverbial (núcleo advérbio). Os sintagmas nominais são as expressões lingüísticas utilizadas para referenciar entidades em um discurso.

No caso do sintagma nominal, o núcleo pode configurar-se em nome comum/próprio ou pronome: pessoal, demonstrativo, indefinido, possessivo entre outros. O sintagma nominal pode apresentar ainda determinantes e/ou modificadores. Os determinantes antecedem o núcleo, podendo ser artigos definidos (o, a, os e as), indefinidos (um, uma, uns, umas), pronomes possessivos (meu, minha, seu, teu etc) entre outros. Os modificadores antecedem

ou sucedem o núcleo. Por exemplo, em "*as acusações*", observa-se um sintagma nominal constituído por um determinante, na forma de artigo definido (*as*), e um núcleo, na forma de substantivo (*acusações*). Já em "*as acusações contra policiais*", há o mesmo sintagma nominal do exemplo anterior, só que agora modificado pelo sintagma preposicional (*contra policiais*).

Dentre as várias configurações de sintagmas nominais, observaremos as descrições definidas, foco de nosso trabalho. Muitos lingüistas chamam a atenção para uma estrutura bastante complexa das descrições definidas em Língua Portuguesa (KOCH, 2000; FÁVERO, 1997; VILELA, KOCH, 2001; KOCH, TRAVAGLIA, 2002; MACAMBIRA, 1990; HAUSSER, 1999; JURAFSKY, MARTIN, 2000).

Dessa forma, as descrições definidas podem ser constituídas por uma extensa e variável seqüência de termos (BONINI, 2002, KOCH, TRAVAGLIA, 1996; ZUMTHOR, 2000). As possibilidades de estruturações de descrições definidas que acreditamos serem importantes à distinção de expressões novas no discurso e outra serão vistas na seção 2.4; porém, antes, são apresentados conceitos sobre correferência e anáfora.

2.2 Correferência e Anáfora

Os conceitos de correferência e anáfora são similares, com pequenas diferenças. Expressões correferentes fazem referência à mesma entidade, enquanto expressões anafóricas podem retomar uma referência anterior (nesse caso, correferentes) ou podem ativar um novo referente cuja interpretação é dependente de outras expressões referenciais anteriormente presentes do texto (nesse caso, não correferentes).

Geralmente uma expressão correferente é anafórica, mas nem sempre uma expressão anafórica é correferente, veja o exemplo:

"O Eurocenter oferece cursos de Japonês na bela cidade de Kanazawa. Os cursos têm quatro semanas de duração. As aulas do nível avançado incluem refeições típicas e passeios a pontos turísticos".

No exemplo acima, a expressão "*Os cursos*" retoma a expressão anterior "*cursos de Japonês"*, ou seja, as duas expressões referenciais fazem menção à mesma entidade, portanto duas expressões correferenciais e anafóricas. Porém, expressões anafóricas não precisam ser necessariamente correferentes, por exemplo, a expressão "*As aulas do nível avançado*" não é correferente a nenhum termo anterior, mas apresenta parte do seu significado apoiado na expressão "*cursos de Japonês*", portanto trata-se de uma expressão não correferente e anafórica.

Na maioria das vezes, uma expressão anafórica ou correferente manifesta-se como um pronome ou uma descrição definida[1] ou demonstrativa[2] (FÁVERO, KOCH, 1994; KOCK, 2000). Porém, ao contrário dos pronomes e das descrições demonstrativas, estudos de corpus da Língua Inglesa (VIEIRA, 1998) e Portuguesa (SALMON-ALT, VIEIRA, 2002; VIEIRA; SALMON-ALT; SCHANG, 2002; VIEIRA et al., 2002; VIEIRA; GASPERIN; GOULART, 2003) mostram que 50% das descrições definidas não possuem um antecedente textual. Esse dado chama a atenção para a necessidade da elaboração de uma estratégia de resolução de anáforas ou correferência que seja composta por uma etapa que identifique as expressões novas no discurso a fim de que as operações de verificação de antecedentes ocorram exclusivamente nas expressões anafóricas ou correferentes.

Com o objetivo de contribuir para essa etapa de classificação, foram analisadas características que predominantemente se manifestariam em expressões novas no discurso (seção 3.7). Porém, antes de prosseguir, para uma boa compreensão deste estudo de

[1] Grupo de palavras que inicia por artigo definido e possui núcleo nome (e.g. *o parecer*, *a resolução final*).
[2] Grupo de palavras que inicia por pronome demonstrativo e possui núcleo nome (e.g. *essa resolução*).

características; as possíveis classificações das descrições definidas serão apresentadas a seguir.

2.3 **Classificação das descrições definidas**

Em Vieira (1998), encontra-se uma divisão das descrições definidas em quatro categorias, dependendo da forma que estão relacionadas com os seus antecedentes:

1. *Anafóricas Diretas*: são antecedidas por uma expressão que possui o mesmo nome-núcleo e refere-se à mesma entidade no discurso. Por exemplo:

"*As listas apontam quase todas as divisões e departamentos da Polícia Civil. Alguns delegados da Polícia Federal também são citados nas listas*".

2. *Anafóricas Indiretas*: são antecedidas por uma expressão que não têm o mesmo nome-núcleo do seu antecedente, mas referem-se à mesma entidade já introduzida no discurso. Assim, o núcleo pode ser um sinônimo do antecedente ou mesmo uma elipse. Por exemplo:

"A Folha de São Paulo apresentou as listas apreendidas na operação contra o crime organizado. O jornal tentou ouvir o delegado encarregado".

3. *Anafóricas Associativas*: possuem um antecedente textual não correferente. Assim, a descrição definida tem seu significado ancorado em uma outra entidade. Por exemplo:

"*A Folha de São Paulo apresentou as listas apreendidas na operação contra o crime organizado. O jornal tentou ouvir o delegado encarregado*".

4. *Novas no discurso*: são aquelas que introduzem um novo referente no texto e não possuem uma âncora para se apoiar semanticamente. Por exemplo:

"*O quilômetro 430 da rodovia Assis Chateau Briand*" *ontem foi cenário da campanha de segurança no trânsito*".

Cabe ressaltar que as expressões anafóricas diretas e indiretas são expressões correferentes, pois se referem à mesma entidade. Já as expressões anafóricas associativas não são correferentes, pois apesar de possuírem uma relação semântica com seus antecedentes, o antecedente não referencia a mesma entidade. As expressões novas no discurso não são correferentes, pois é a primeira manifestação das expressões no discurso.

A seguir, apresentamos o estudo de características das descrições definidas novas no discurso.

2.4 Estudo das características das descrições definidas

Na literatura encontramos várias propostas que levam em consideração a estrutura sintática; aspectos lexicais; semânticos; posicionais; de contexto; entre outras características para classificar as descrições definidas, conforme apresentado em (MCCARTHY, LEHNERT, 1995; BEAN, RILOFF, 1999; CARDIE,WAGSTAFF,1999; VIEIRA, POESIO, 2000; SOON; NG; LIM, 2001; MULLER; RAPP; STRUBE, 2002; NG, CARDIE, 2002a; NG, CARDIE, 2002b; URYUPINA, 2003; POESIO et al., 2005).

Um resumo destes trabalhos é apresentado na seção 2.6. Muitos desses trabalhos referem-se à estrutura do sintagma como característica relevante. Um estudo de corpus foi realizado para verificar a presença destas e de outras características (seção 3.7).

Quanto ao aspecto estrutural, é importante discriminar características que são comuns, e por isso presentes em todas as descrições definidas, das que são adicionais, e que, em alguns casos, podem sinalizar expressões novas no discurso. Nesse sentido, nosso olhar concentra-se em dois blocos distintos de elementos estruturais das descrições definidas: as partes fundamentais e os complementos.

2.4.1 **Partes fundamentais das descrições definidas**

Uma descrição definida é fundamentalmente composta por duas classes de palavras:

- Artigo definido: palavra que acompanha o substantivo, determinando-o de forma definida.

Na Língua Portuguesa os artigos definidos são: o, a, os e as.

- Nome (substantivo): palavra que designa os seres/objetos reais ou imaginários. O nome pode ser comum (referência genérica a um ser/objeto) ou próprio (referência específica a um ser/objeto, identificando este entre todos os outros seres/objetos de uma espécie).

O núcleo nome pode ser formado por um ou mais nomes. Nos casos da presença na estrutura sintática de um único nome, ele é denominado *simples;* caso contrário, ele será denominado *composto*.

Além disso, na composição das descrições definidas, com exceção dos artigos definidos, que assumem a posição primeira no sintagma, os nomes, apesar de centrais, não ocupam uma posição fixa, pois, para referenciação, por vezes, vários elementos complementadores são necessários, ora precedendo os nomes ora antecedendo os nomes.

2.4.2 **Os complementos das descrições definidas**

Os elementos complementares das descrições definidas também são mencionados na literatura como elementos modificadores. Na construção das descrições definidas, elementos que se somam aos nomes são:

- Aposto: é um ou mais termos que se referem a um substantivo ou pronome explicando-o.

O aposto pode localizar-se entre vírgulas ou travessões ou vir depois de dois-pontos3. Para exemplificar, segue a descrição definida:

[3] No corpus é comum a presença desse tipo de construção sem o uso de travessões ou entre vírgulas.

"*O bicheiro*, *Castor de Andrade,"*.

- Sintagma preposicional: sintagma que possui como núcleo uma preposição, palavra invariável que liga/relaciona dois termos (palavras e orações), indicando origem, posse, finalidade, meio, causa etc.

Nessa relação, um termo vai explicar ou completar o sentido do outro. Para exemplificar, segue a descrição definida:

"*A data da reforma"*.

- Sintagma adjetival: sintagma que possui como núcleo um adjetivo (palavra que especifica e caracteriza seres/objetos atribuindo-lhes estados ou qualidades).

O sintagma adjetival também pode apresentar intensificador (palavra que modifica um verbo, um adjetivo ou outro advérbio, indicando uma circunstância). Para exemplificar, seguem as descrições definidas:

"*A literatura infantil"*.

"*As democracias mais fortes"*.

- Cláusula relativa: são orações que funcionam como sintagmas adjetivais e apresentam-se encaixadas na posição de modificador do nome. Por exemplo:

"*O sistema que fiscalizava o Inamps"*.

Podemos dizer, então, que as descrições definidas possuem um conjunto de configurações conforme Tab. 2.1, seguindo a legenda abaixo:

Legenda:

AD: artigo definido;
N: nome;
SA: sintagma adjetival;
SP: sintagma preposicional;
REL: cláusula relativa;
PRONP: pronome possessivo;
INT: intensificador;
APO: aposto.

Tabela 2.1 Configurações das descrições definidas.

Configurações	Exemplos
AD + N	*A humanidade.*
AD + N + SA	*A vista ocidental.*
AD + SA + N	*Os grandes clientes.*
AD + SA + N + SA	*As antigas casas medievais.*
AD + NOME + SP	*O artefato de lentes.*
AD + PRONP + N + SP	*O meu amigo de infância.*
AD + PRONP + N + SA	*A nossa proposta ambiental.*
AD + PRONP + SA + N + SA	*As nossas grandes façanhas administrativas.*
AD + SA + N + SP	*A antiga cidade de Kanazawa.*
AD + N + INT + SA + SP	*Os momentos mais difíceis de minha carreira.*
AD + N + REL	*O comunicado que deve ser assinado pelos jornalistas.*
AD + SA + N + SP + REL	*As grandes deficiências da gestão financeira que recentemente provocou pela primeira vez na história do clube atraso no pagamento dos jogadores.*
AD + N + SP + REL	*O texto do comunicado que deve ser assinado hoje.*
AD + N + APO	*O ex-deputado, Agnaldo Timóteo,*
AD + N + SP + APO	*O diretor do Departamento de Polícia do Interior, delegado Mário Covas,*
AD + N + REL + OPO	*O delegado que é acusado pelo departamento, Élson Campelo,*
AD + N + INT + SA + REL + APO	*O deputado muito disposto que iniciara as investigações, Emir Laranjeira,*

2.5 Árvores de Decisão

O Aprendizado de Máquina (AM) é uma sub-área de pesquisa em Inteligência Artificial (IA) relacionada à capacidade de aprender ser essencial para um comportamento inteligente. O objetivo principal de AM é encontrar métodos computacionais baseados em operações lógicas ou binárias capazes de aprender uma tarefa a partir de um conjunto de exemplos. Neste sentido, o AM estuda métodos computacionais para adquirir novos conhecimentos, novas habilidades e novos meios de organizar o conhecimento já existente (MITCHELL, 1997). O estudo de técnicas de aprendizado baseado em computador também pode fornecer um melhor entendimento do próprio raciocínio.

A indução é a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos, a qual se caracteriza pelo raciocínio a partir de um conceito específico e generalizado, ou seja, da parte para o todo. Sendo assim, na indução um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados (MONARD, BARANAUSKAS, 2003).

O aprendizado indutivo pode ser dividido em supervisionado e não-supervisionado. No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. Um rótulo descreve o fenômeno de interesse, isto é, o conceito-meta que se deseja aprender para fazer previsões a respeito.

No aprendizado supervisionado, geralmente cada exemplo é descrito por um vetor de valores de características ou de aspectos do exemplo, esses conhecidos como atributos, e pelo rótulo da classe associada. Tem-se como objetivo do algoritmo de indução, construir um classificador que seja capaz de determinar corretamente a classe de novos exemplos que ainda não possuam rótulo da classe. Sendo que, para rótulos de classe discretos, esse problema é conhecido como Classificação e para valores contínuos como Regressão.

Dentro da área de Aprendizado de Máquina foram propostos vários paradigmas capazes de aprender a partir de um conjunto de exemplos, constituindo-se de paradigmas simbólico, estatístico, conexionista, evolutivo, baseado em casos (MONARD, BARANAUSKAS, 2003). No contexto desse trabalho, será abordado o paradigma simbólico, no qual os sistemas objetivam aprender constituindo representações simbólicas de um conceito por meio da análise de exemplos e contra-exemplos desse conceito. Dentre as representações simbólicas, será utilizada a representação por árvores de decisão para a classificação de entidades novas no discurso. Adotamos árvores de decisão, pois são geralmente utilizadas para esse tipo de tarefa, conforme podemos observar na literatura da

área. Além disso, o método é adequado para a análise da relevância das características que é um dos nossos objetivos principais.

Uma árvore de decisão é capaz de prever a classe de um exemplo baseada em decisões realizadas sobre os atributos que descrevem esse exemplo (AMADO, 2001). Neste contexto, uma árvore de decisão é uma forma simples de representação, que classifica exemplos de uma base de dados em um número finito de classes, tendo a seguinte estrutura:

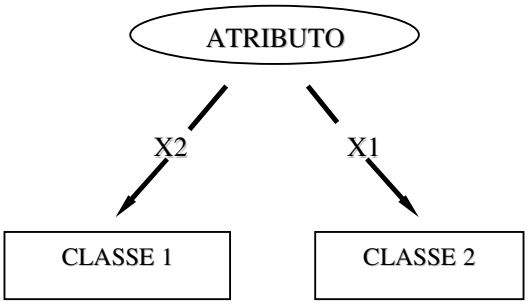

FIGURA 2.1 Estrutura da Árvore de Decisão.

Na FIGURA 2.1 cada nodo representa um atributo da base de dados. Cada galho representa um valor do atributo. Cada nodo folha representa uma classe. Além dessa estrutura, as árvores de decisão possuem algumas características, tais como:

- Representam uma série de perguntas em relação aos atributos do domínio;
- Um objeto é classificado seguindo o caminho do nodo raiz até o nodo folha, enquanto as suas características satisfazem as ligações.

As árvores de decisão são baseadas no uso de um algoritmo de particionamento recursivo, tais como ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984). Esses realizam uma busca em um espaço de hipóteses completo e possuem um viés indutivo em direção a árvores de tamanho reduzido, baseando-se na abordagem: "*dividir para conquistar*".

Nesses algoritmos, a indução da árvore baseia-se na divisão recursiva do conjunto de exemplos de treinamento em subconjuntos mais representativos, ou seja, dividir o conjunto de

treinamento até todos os subconjuntos conterem exemplos que pertencem a apenas uma classe. As divisões do conjunto de treinamento são realizadas com base nos valores possíveis de um dos atributos que descrevem os exemplos. A cada divisão é associado um nodo na árvore que representa o teste realizado aos valores do atributo que deu origem à divisão. Quando todos os exemplos de um conjunto pertencem à mesma classe, esse é associado a uma folha na árvore que representa a classe dos exemplos.

O tamanho e a precisão das árvores de decisão construídas por este processo depende da escolha dos atributos utilizados para a divisão do conjunto de treinamento, sendo o critério mais utilizado, para escolher o atributo que particiona o conjunto de exemplos em cada iteração, a métrica de ganho de informação[4]. Um bom critério de escolha de atributos é fundamental, uma vez que uma má escolha de um atributo pode fragmentar o conjunto de treinamento e reduzir a precisão.

Uma outra questão a ser considerada na construção das árvores de decisão é a simplificação, uma vez que o método descrito para a construção das árvores de decisão resulta em árvores perfeitamente ajustadas ao conjunto de exemplos utilizados na sua construção, ou seja, ocorre uma superestimativa dos dados (*overfitting*[5]). Para tentar solucionar o problema de superajuste dos dados, utilizam-se dois métodos de simplificação das árvores de decisão:

- O primeiro método consiste em utilizar um determinado critério para verificar se, antes de realizar a divisão dos exemplos num nodo, a divisão é relevante para a classificação final. Caso não o seja, esta não é realizada e a exploração do galho da árvore é terminada.

[4] O ganho de informação é uma medida que indica a redução esperada na entropia de um conjunto de dados, causada pelo particionamento dos exemplos em relação a um dado atributo.
[5] *Overfitting* ocorre quando ao induzir, a partir de exemplos disponíveis, a hipótese é muito específica para o conjunto de treinamento utilizado, ou seja, a hipótese ajusta-se em excesso ao conjunto de treinamento.

- O segundo método, muitas vezes denominado poda, é aplicado após a construção da árvore. Cada nodo da árvore é visitado e os galhos que não contribuírem significativamente para a classificação são simplificados.

As árvores de decisão podem ser representadas também por conjuntos de regras "*if-then*", por serem estas mais legíveis. Cada regra representa um possível caminho a ser percorrido desde a raiz até uma folha, onde o resultado da classificação é especificado.

Nesse trabalho, as árvores de decisão serão aplicadas na investigação de quais atributos são mais relevantes para a classificação das descrições definidas como novas no discurso, com base no corpus anotado manualmente.

Uma vez apresentados os conceitos sobre árvores de decisão, na seqüência, apresentaremos alguns trabalhos relacionados que utilizam esses conceitos e contribuíram para a dissertação.

2.6 Trabalhos Relacionados

Nesta seção são apresentados alguns trabalhos relacionados sobre resolução de correferência de sintagmas nominais, focalizando as características tanto sintáticas como semânticas dos sintagmas nominais, utilizadas no processo de resolução de anáforas.

Muitas aplicações em PLN requerem a resolução de anáforas de sintagmas nominais, sendo necessários meios para determinar quais sintagmas nominais em um texto ou em um diálogo referem-se à mesma entidade real do mundo.

A maioria dos algoritmos que tratam resolução de anáforas de sintagmas nominais combina características sintáticas e semânticas, com o intuito de desenvolver sistemas robustos. A seguir, serão descritos alguns trabalhos relacionados à resolução de anáforas e uma visão geral das características que eles apresentam.

Aone e Bennett (1995) descrevem um sistema de resolução de anáfora que abrange tipos de anáforas (descrições definidas, pronominais etc.) que podem ser definidos a partir da necessidade do usuário. Para a realização dos experimentos, um conjunto de artigos jornalísticos em japonês foi usado como exemplos de treinamento para o algoritmo de aprendizado de máquina de árvores de decisão C4.5 (QUINLAN, 1993). Para isso, Aone e Bennett utilizaram 66 características, dentre as quais estão as características morfológicas, sintáticas, semânticas e posicionais. Essas características podem ser características unárias[6] ou características binárias[7].

Em Mccarthy e Lehnert (1995) a resolução de anáforas é apresentada como um problema dos Sistemas de Extração de Informação, que necessitam identificar informações de interesse em uma coleção de textos, onde as entidades envolvidas são referenciadas em lugares diferentes e em caminhos diferentes. Esse problema pode ser reformulado como um problema de classificação: dado dois referentes, eles apontam para os mesmos objetos ou para objetos diferentes. Neste sentido, Mccarthy e Lehnert apresentaram uma nova abordagem para resolução de correferência (Sistema RESOLVE): a construção de árvores de decisão que classificam pares de sintagmas nominais como correferentes ou não correferentes. Mccarthy e Lehnert relatam o uso de oito características para essa tarefa de classificação de correferência, dentre essas características temos a presença de um nome próprio e a presença de um sintagma nominal comum em um par de referentes.

Bean e Riloff (1999) desenvolveram um algoritmo baseado em corpus para identificar automaticamente sintagmas nominais definidos não-anafóricos que possuem potencial para melhorar a eficiência e a precisão de sistemas de resolução de correferência. Bean e Riloff classificaram os sintagmas nominais definidos utilizando a seguinte taxonomia: sintagmas nominais referentes que possuem um referente anterior nos textos (antecedente), e sintagmas

[6] Característica de uma anáfora ou de um antecedente.
[7] Característica que diz respeito às relações entre os pares de anáforas e de antecedentes.

nominais existenciais[8] que não possuem um referente anterior nos textos (antecedente). Após, os sintagmas nominais foram classificados em duas categorias: independentes[9] e associativos[10]. O objetivo de Bean e Riloff é construir um sistema para identificar os sintagmas nominais existenciais independentes automaticamente. Para isso, observaram-se as características de "existencialismo" dos sintagmas nominais definidos a partir da sintaxe e da semântica. Desta forma, foi construído um conjunto de heurísticas sintáticas que procuram pistas (*cues*) estruturais de pré-modificadores restritivos e pós-modificadores restritivos.

Em Cardie e Wagstaff (1999) um novo algoritmo não supervisionado para resolução de correferência de sintagmas nominais é apresentado. A resolução de correferência é tratada como uma tarefa de agrupamento[11]. Cardie e Wagstaff utilizaram onze características para particionar os sintagmas nominais simples[12], que foram obtidas automaticamente sem qualquer etiquetagem manual. Dentre essas características verificou-se, por exemplo, o tipo de artigo (definido, indefinido, nenhum); o tipo de pronome (possessivo, ambíguo, nominal); o gênero (masculino, feminino, neutro, outro); entre outras.

Em Vieira e Poesio (2000) é apresentado um sistema para o processamento de descrições definidas independente de domínio, que está fundamentado em experimentos baseados em corpus. O sistema foi implementado e testado com idéias de diferentes usos das descrições definidas e observou-se a predominância de descrições novas no discurso em corpus jornalísticos na Língua Inglesa. Vieira e Poesio detalham um estudo sobre as descrições definidas na Língua Inglesa, utilizando a classificação dos usos de descrições

[8] Um sintagma nominal definido é existencial, quando especifica completamente uma representação cognitiva da entidade na mente do leitor, por exemplo, "O F.B.I.".
[9] Sintagmas nominais existenciais independentes são entendidos isoladamente pelo leitor, sem a necessidade de um contexto.
[10] Sintagmas nominais existenciais associativos são inerentemente associados a um evento, ação, objeto ou outro contexto, para o entendimento do leitor.
[11] Agrupamento é um método de descoberta de conhecimento utilizado para identificar co-relacionamentos e associações entre objetos, facilitando assim a identificação de classes.
[12] Sintagmas nominais simples que não contém nenhum outro sintagma nominal menor dentro dele.

definidas[13] abordadas nos experimentos de Vieira (1998). E assim, foram desenvolvidos três diferentes conjuntos de heurísticas para:

1. Resolver descrições diretamente anafóricas;

2. Identificar descrições novas no discurso;

3. Identificar uma âncora da descrição associativa e a relação semântica entre a descrição associativa e a sua âncora.

Soon; Ng; Lim (2001) apresenta uma abordagem de aprendizado para resolução de correferência de sintagmas nominais em textos do MUC-6[14] (MUC-6, 1995) e do MUC-7 (MUC-7, 1997). A abordagem utiliza, para o aprendizado, um corpus pequeno anotado e a sua tarefa é determinar relações de correferência entre elementos textuais, como por exemplo, sintagmas nominais definidos, sintagmas nominais demonstrativos, nomes próprios, apostos. Soon; Ng; Lim relatam o uso de doze características para verificar se duas entidades são ou não correferentes (antecedente em potencial e anafórica). Dentre as características temos a verificação se o antecedente em potencial e o anafórico são pronomes (reflexivos, pessoais, possessivos); verificação se o antecedente em potencial e o anafórico são nomes próprios; identificação do tipo de sintagma nominal (definido, demonstrativo); verificação de construções de aposto; verificação de flexão de número (plural, singular) e de gênero (feminino e masculino); entre outras.

Ng e Cardie (2002a) relatam um método de aprendizado supervisionado para a identificação de sintagmas nominais anafóricos e não anafóricos e mostram como podem ser incorporadas tais informações em um sistema de resolução de correferência. Para isso, Ng e Cardie construíram um classificador para a determinação da anaforicidade, utilizando o sistema de indução de árvore de decisão C4.5 (QUINLAN,1993), onde cada exemplo de treinamento representa um único sintagma nominal e consiste de 37 características que são

[13] As descrições definidas estão divididas em três categorias (anáforas diretas, novas no discurso e associativas), dependendo da forma que estão relacionadas com os seus antecedentes.
[14] MUC: Conferências organizadas em forma de competição para apresentação de sistemas.

potencialmente úteis para distinguir entre sintagmas nominais anafóricos e não anafóricos. Lingüisticamente, essas características podem ser divididas em quatro grupos: léxico, gramatical, semântico e posicional.

Em Ng e Cardie (2002b) apresenta-se um sistema de correferência de sintagmas nominais que estende o trabalho de Soon; Ng; Lim (2001) e produz melhores resultados para o conjunto de dados de resolução de correferência do MUC-6 (MUC-6, 1995) e do MUC-7 (MUC-7, 1997). Ng e Cardie, então, estenderam o conjunto de características de Soon; Ng; Lim (2001) de 12 características para um conjunto mais amplo de 53 características, sendo características léxicas, semânticas, e baseadas em conhecimento; além de 26 características gramaticais que incluem uma variedade de restrições lingüísticas e preferências.

Em Muller; Rapp; Strube (2002) relata-se a realização de alguns experimentos com o algoritmo de Aprendizado de Máquina Supervisionado Co-treinamento[15]. Com o intuito de verificar se o algoritmo de Co-Treinamento pode reduzir significativamente a quantidade de trabalho manual de etiquetagem e ainda produzir um classificador com um desempenho aceitável. Para isso, utilizou-se um corpus com 250 textos em alemão sobre locais turísticos, eventos históricos e pessoas em Heidelberg. As características utilizadas por Muller; Rapp; Strube foram consideradas como independentes do domínio, distinguiu-se entre as características atribuídas a sintagmas nominais, tais como função gramatical do antecedente e do anafórico, e as características atribuídas à relação de correferência potencial, tais como a distância entre o anafórico e o antecedente em palavras e em sentenças, totalizando um conjunto de 17 características.

Strube; Rapp; Muller (2002) mostram uma abordagem de árvore de decisão[16] que utiliza um conjunto de características usadas em trabalhos prévios em experimentos de resolução de correferência (SOON; NG; LIM, 2001; CARDIE, WAGSTAFF, 1999;

[15] É um algoritmo que utiliza exemplos de treinamento não etiquetados, além dos etiquetados para o aprendizado do classificador.
[16] Algoritmo de classificação de árvores de decisão *j48*, o qual é uma re-implementação em Java do C4.5.

MCCARTHY, LEHNERT, 1995) e características adicionais independentes de domínio baseadas na mínima distância de edição (*MEC*) entre *strings*. Nesse estudo, analisaram-se o desempenho desse conjunto de características para as diferentes formas de expressões anafóricas, encontrando bons resultados para pronomes, resultados moderados para nomes próprios e resultados pobres para sintagmas nominais definidos. Strube; Rapp; Muller analisaram também a influência das características baseadas na distância mínima de edição *(MEC*) entre o anafórico e o antecedente na resolução de referência, que são computadas pelo número de substituições, inserções, deleções e pelo comprimento do antecedente em potencial ou do anafórico.

Uryupina (2003) desenvolveu um sistema para identificação automática de entidades novas no discurso e únicas, com base no discurso e no ouvinte utilizando o algoritmo de aprendizado de máquina RIPPER (COHEN, 1995). Nesse sentido, classificaram-se as entidades em novas no discurso[17] e velhas no discurso, e as expressões em referidas unicamente[18] e referidas não unicamente. Para isso, utilizou-se um corpus pequeno de treinamento do MUC-7 (MUC-7, 1997), além de alguns dados da Internet. Uryupina usou trinta e duas características para a realização dos experimentos, que estão divididas em características sintáticas, como a identificação de apostos; características de contexto, como o cálculo da distância entre um determinado sintagma nominal e o seu antecedente em potencial de mesmo núcleo; e características de probabilidade definida, onde para cada expressão buscou-se na internet[19] o número de páginas que contenham tal expressão.

Poesio et al. (2005) reexaminaram a literatura referente à resolução de anáforas (VIEIRA; POESIO, 2000; NG; CARDIE, 2002a; URYUPINA, 2003) e propuseram um algoritmo revisado que incorpora um novo conjunto de características para a descoberta de

[17] Entidades novas no discurso quando se refere à um objeto ou pessoa não mencionada previamente no discurso.
[18] Expressões são únicas quando especificam completamente o seu referente, sendo interpretada sem qualquer contexto.
[19] Site de busca do Alta Vista, disponível em: http://www.altavista.com/

descrições definidas novas no discurso. Este algoritmo segue dois passos: primeiro, é executado o algoritmo de resolução de anáforas diretas de Vieira e Poesio (2000), além de outras características de detecção das expressões novas no discurso da literatura, que serão as características de entrada para o classificador. Segundo, um classificador baseado em árvores de decisão (implementação do C4.5 incluindo a biblioteca da Ferramenta Weka 3.4) é utilizado para classificar as descrições definidas como anafóricas (caso tenha sido encontrado um antecedente no primeiro passo) ou novas no discurso. As características de entrada das descrições definidas baseiam-se no reconhecimento de predicativos (construções copulares, aposto); nomes próprios; funcionalidade (superlativo); cláusulas relativas e posição no texto.

Os trabalhos relacionados com resolução de anáforas apresentam propostas para o tratamento de diferentes tipos de anáforas com diferentes conjuntos de características analisadas. Cabe ressaltar que, as características utilizadas para a resolução das descrições definidas nestes trabalhos serão identificadas e utilizadas na construção das árvores de decisão, na busca da melhor combinação destas características para a classificação das descrições definidas novas no discurso, com base no corpus anotado manualmente.

Cabe ressaltar que, todos os trabalhos citados acima se referem a outras línguas diferentes da Língua Portuguesa. Para a Língua Portuguesa existem estudos de corpora sobre correferência, mas esses trabalhos não implementam resolução ou classificação automática (SALMON-ALT, VIEIRA, 2002; VIEIRA; SALMON-ALT; SCHANG, 2002; VIEIRA et al., 2003; VIEIRA; GASPERIN; GOULART, 2003).

2.7 Considerações Finais

Neste capítulo foram apresentados conceitos importantes de lingüística e computação. Com base nestes conceitos, é possível perceber a importância da identificação de expressões referenciais que não possuem antecedentes.

Na literatura encontramos vários trabalhos que apresentam algoritmos que tratam a resolução de anáforas de sintagmas nominais e partem da investigação de suas características tanto sintáticas quanto semânticas na busca de um bom classificador. Entre os algoritmos de classificação, destacam-se os de árvores de decisão (seção 2.5) por serem mais adequados para a análise da relevância de características. Partimos de estudos realizados em outras línguas, especialmente na Língua Inglesa, com o objetivo de investigar quais características são mais significativas para o processo de classificação das descrições definidas. Algumas das características analisadas nos trabalhos relacionados são ilustradas na Tab. 2.2.

Árvores de decisão são importantes para a investigação de atributos relevantes para a classificação das descrições definidas.

Tabela 2.2 Algumas características de trabalhos relacionados.

	Mccarthy; Lehnert	**Bean; Riloff**	**Cardie; Wagstaff**	**Vieira; Poesio**	**Soon; Ng; Lim**	**Ng; Cardie**	**Poesio et al**
CARACTERÍSTICAS		SP[20]		SP[21]			SP[21]
		APO[20]	APO	APO[20]		APO	APO[20]
	APO_NP[22]	APO_NP[23]	APO_NP[22]	APO_NP[23]	APO_NP[22]		APO_NP[23]
		REL[20]		REL[21]		REL	REL[21]
	NP_COM[22]		NP_COM[22]	NP_COM[24]	NP_COM[22]	NP_COM	NP_COM[24]
				PRE_ADJ[25]		PRE_ADJ	
		PRE_NUM				PRE_NUM	
				SUP[25]		SUP	SUP
				COP			COP
	PRI_SENT		PRI_SENT		PRI_SENT		PRI_SENT
		SEM_ANT		SEM_ANT		SEM_ANT	SEM_ANT

Legenda:

SP: sintagma preposicional;
APO: aposto;
APO_NP: nome próprio com função de aposto;
NP_COM: núcleo nome próprio composto;
PRE_ADJ: pré modificador adjetivo;
PRE_NUM: pré modificador número;
SEM_ANT: verificação de um antecedente.
SUP: superlativo;
COP: construção copular;
PRI_SENT: primeira sentença;
REL: cláusula relativa;

[20] Características de um mesmo grupo.
[21] Características de um mesmo grupo sendo analisadas como pós modificadores restritivos.
[22] Analisam somente a presença de um nome próprio.
[23] Característica sendo analisada como modificador nome próprio.
[24] Núcleo nome próprio simples.
[25] Características mais restritivas (por exemplo, listas de predicados especiais).

3. Materiais e Métodos

Este capítulo apresenta o corpus e as ferramentas utilizadas nos experimentos: o analisador sintático PALAVRAS, usado na análise gramatical dos textos; a ferramenta Xtractor que converte a análise para o código XML; a ferramenta MMAX, usada na marcação manual do corpus; a ferramenta ART que trata a resolução de anáforas; a ferramenta Weka, constituída de uma coleção de algoritmos de aprendizado de máquina. Após, os grupos de características usados para a classificação e o processo de identificação automática das descrições definidas são mostrados. Por fim, a metodologia proposta para o aprendizado é descrita, além da implementação das árvores de decisão no ambiente ART.

3.1 Descrição do Corpus

Neste trabalho, foram utilizados dois corpora. O primeiro (corpus 1) constitui-se de um extrato do corpus NILC[26], formado por um conjunto de 24 textos jornalísticos da Folha de São Paulo, escritos em português do Brasil. Cada documento é um arquivo texto (formato ASCII) com tamanho entre 1 Kbytes e 6 Kbytes, com um mínimo de 186 palavras e um máximo de 1089 palavras, totalizando 11042 palavras. O corpus contém 2319 sintagmas nominais, sendo que 1411 são descrições definidas. Estas informações são apresentadas na Tab 3.1.

TABELA 3.1 Informações sobre o corpus 1.

Corpus	Nº Textos	Tamanho	Nº total de Palavras	Nº Sintagmas Nominais	Nº Descrições Definidas
corpus 1	24	de 1 à 6 Kb	11042	2319	1411

[26] Núcleo Interinstitucional de Lingüística Computacional. Disponível em http://www.nilc.icmp.usp.br/nilc

Além deste, outro corpus (corpus 2) será utilizado para validação dos experimentos, o qual constitui-se de um extrato do corpus Público (SANTOS, 2000) formado por 4 textos retirados do jornal Público, escritos no português europeu. Cada documento é um arquivo texto (formato ASCII) com tamanho entre 2 Kbytes e 9 Kbytes, com um mínimo de 201 palavras e um máximo de 1356 palavras, totalizando 3627 palavras. O corpus contém 777 sintagmas nominais, sendo que 483 são descrições definidas. Estas informações são apresentadas na Tab 3.2.

TABELA 3.2 Informações sobre o corpus 2.

Corpus	Nº Textos	Tamanho	Nº total de Palavras	Nº Sintagmas Nominais	Nº Descrições Definidas
corpus 2	4	de 2 à 9 Kb	3627	777	483

Os corpora usados foram anotados com informações de correferência. A ferramenta MMAX foi utilizada para a realização da anotação manual e será descrita na seção 3.3. Os seguintes estudos de corpora estão relacionados com este trabalho e foram desenvolvidos com a ajuda da ferramenta MMAX: (SALMON-ALT, VIEIRA, 2002; VIEIRA; SALMON-ALT; SCHANG, 2002; VIEIRA et al., 2003; VIEIRA; GASPERIN; GOULART, 2003).

A anotação manual do corpus 1 foi realizada inicialmente por dois anotadores (Cassiano Ricardo Haag e Terezinha Margarete da Silva) em quatro etapas. Em um primeiro momento, foram anotadas as descrições definidas, considerando-se que uma descrição definida pode conter outras descrições definidas, por exemplo, *"A lista do banqueiro do jogo do bicho", "o banqueiro do jogo do bicho", "o jogo do bicho"*. Em um segundo momento, foi analisada a correferência das expressões, sendo estas classificadas como correferentes e não correferentes. Em um terceiro momento, as descrições definidas correferentes foram classificadas como anafóricas diretas e anafóricas indiretas, sendo que para estas foram apontados os respectivos antecedentes. Em uma etapa final, as descrições definidas não correferentes, foram classificadas em novas no discurso e anafóricas associativas, sendo que,

para as anafóricas associativas foram apontadas as expressões em que estas estão ancoradas semanticamente (antecedentes). Essa anotação foi revisada pelo anotador Jorge Cesar Barbosa Coelho.

No corpus 2, a anotação manual foi realizada pelo anotador Jorge Cesar Barbosa Coelho, seguindo as mesmas etapas do corpus 1.

Na próxima seção é apresentado o analisador sintático usado na análise gramatical dos textos e a ferramenta Xtractor que converte a análise para o código XML.

3.2 PALAVRAS e Xtractor

Em muitas aplicações de PLN, é necessário utilizar um analisador sintático, que trabalha em nível de sentença ou sintagma e reconhece uma seqüência de palavras como constituintes de uma frase da língua construindo uma árvore de derivação, que explicita as relações entre as palavras que compõem a sentença.

Neste trabalho foi utilizado o analisador sintático PALAVRAS (BICK, 2000), uma ferramenta robusta utilizada para a análise sintática do português. A partir da saída deste analisador sintático, a ferramenta Xtractor[27] (GASPERIN et al., 2003) gera três arquivos XML (*eXtensible Markup Language*).

O primeiro é o arquivo básico de palavras (*words)*; o segundo é o arquivo com as categorias morfossintáticas (*POS – Part of Speech*) das palavras do corpus; e por fim o terceiro é o arquivo com as estruturas sintáticas das sentenças (*chunks)*. Um *chunk* pode possuir sub-elementos *chunks* com informações das sub-estruturas da sentença. Para exemplificar, segue a descrição definida "*O Othon Palace Hotel*" e seus arquivos de *word* (FIGURA 3.1), *POS* (FIGURA 3.2) e *chunks* (FIGURA 3.3).

[27] A Ferramenta Xtractor engloba a análise do corpus a partir do analisador sintático PALAVRAS, o tratamento da saída do analisador sintático, com a geração dos três arquivos XML.

```
<words>
.....
<word id="word_716">o</word>
<word id="word_717">Othon_Palace_Hotel</word>
.....
</words>
```

FIGURA 3.1 Arquivo de *words*.

```
<words>
.....
<word id="word_716">
<art canon="o" gender="M" number="S">
  <secondary_art tag="artd"/>
  <secondary_art tag="-sam"/>
</art>
</word>
<word id="word_717">
<prop canon="Othon_Palace_Hotel" gender="M" number="S"/>
</word>
.....
</words>
```

FIGURA 3.2 Arquivo de *POS*.

```
<text>
<paragraph id= “paragraph_1”>
.....
<sentence id="sentence_24" span="word_698..word_721">
.....
<chunk id="chunk_1018" ext="p" form="np" span="word_716..word_717">
  <chunk id="chunk_1019" ext="n" form="art" span="word_716"/chunk>
  <chunk id="chunk_1020" ext="h" form="prop" span="word_717"/chunk>
</chunk>
.....
</sentence>
.....
</paragraph>
</text>
```

FIGURA 3.3 Arquivo de *chunks*.

Neste trabalho, os atributos dos sub-elementos *chunks* serão utilizados na identificação das características das descrições definidas, que serão mostradas na seção 3.7. Portanto, cabe ressaltar que as informações de interesse dos sub-elementos *chunks* são:

- Atributo *ext*: representa a função sintática do *chunk*, por exemplo, sentença ou enunciado (*ext=sta*); sujeito (*ext=subj*); núcleo (*ext=h*).
- Atributo *form*: representa a forma ou estrutura morfossintática do *chunk*, tais como: cláusula finita (*form=fcl*); sintagma nominal (*form=np*); substantivo (*form=n*).

Na seção que segue é apresentada a ferramenta MMAX que é usada neste estudo para a anotação manual dos corpora utilizados.

3.3 MMAX

Para a anotação manual de corpus, utilizou-se a ferramenta MMAX (*Multi-Modal Annotation in XML*) (MULLER, STRUBE, 2000). Essa ferramenta utiliza o arquivo de *words*, gerado pela ferramenta Xtractor que contém todas as palavras do corpus associadas a um identificador (FIGURA 3.1). Ela também utiliza um segundo arquivo que contém a estrutura do corpus (parágrafos, sentenças, cabeçalhos etc.).

O resultado do processo de anotação no MMAX é um arquivo que contém a anotação de correferência. As marcações são codificadas como elementos *markables*, cujo atributo *span* indica as palavras que formam a expressão, o atributo *pointer* indica o identificador do antecedente; além desses atributos; outros atributos podem ser especificados pelo pesquisador. Em (GOULART; GASPERIN; VIEIRA, 2004) utilizou-se o atributo *classification* que corresponde à classificação anafórica da expressão.

Para exemplificar, segue o trecho de um dos textos do corpus 1:

(1) *"Segundo a Confederação Nacional dos Bispos d**o Brasil** CNBB, a Igreja Católica perde em média, só no Brasil, 600 mil fiéis para outras religiões(i).* [..] *No início do ano, ele enviou ao Brasil 50 discípulos que se hospedaram no Othon Palace Hotel, em São Paulo(ii)".*

Observamos que no exemplo (1) temos duas sentenças: sentença *(i)* corresponde ao *span "word_276..word_297"* e sentença *(ii)* corresponde ao *span "word_698..word_721"* no arquivo da estrutura do corpus (Figura 3.4).

Para uma melhor compreensão do arquivo resultante com as marcações desse estudo, analisemos as marcações das descrições definidas sublinhadas na sentença *(ii)* ilustrada na FIGURA 3.5. A expressão *"o Brasil"* (atributo *span="word_708..word_709"*) possui como

antecedente a expressão *"o Brasil"* que está destacada (em negrito) na sentença *(i)* (atributo *pointer="markable_36"*), a forma de descrição definida (atributo *form="defnp"*), a classificação de anafórica direta (atributo *classification="direct"*), pois o seu antecedente possui o mesmo nome-núcleo da expressão (nome-núcleo "*Brasil"*) e a classificação de correferente porque a expressão analisada e o seu antecedente se referem à mesma entidade (atributo *coreference="coreferent"*). Já a expressão *"o Othon Palace Hotel"* (atributo *span="word_716..word_717*") não possui um antecedente por ser a primeira manifestação da expressão no discurso (atributo *pointer=""*) e possui a forma de descrição definida (atributo *form="defnp"*). Consequentemente, a expressão é classificada como nova no discurso (atributo *classification="discourse_new"*) e não correferente (atributo *coreference="non_coreferent"*).

```
.....
<paragraph>
.....
<sentence id="sentence_10" span="word_276..word_297">
.....
<sentence id="sentence_24" span="word_698..word_721">
.....
</paragraph>
.....
```

FIGURA 3.4 Arquivo da Estrutura.

```
<markable>
.....
<markable id="markable_95" span="word_708..word_709"
pointer="markable_36" form="defnp"
classification="direct" coreference="coreferent"/>
<markable id="markable_96" span="word_716..word_717"
pointer="" form="defnp"
classification="discourse_new" coreference="non_coreferent"/>
.....
</markable>
```

FIGURA 3.5 Arquivo de Marcações.

As ferramentas apresentadas (PALAVRAS, Xtractor, MMAX) são utilizadas no contexto desse trabalho no processo de identificação automática das características das

descrições definidas (seção 3.8) e na metodologia proposta para o processo de aprendizado (seção 3.9).

Na seção que segue são apresentadas a ferramenta ART, que trata a resolução de anáforas, e a ferramenta Weka, formada por um conjunto de algoritmos de AM.

3.4 ART

ART (*Anaphor Resolution Tool*) é uma ferramenta para resolução de expressões anafóricas, entre elas as descrições definidas, onde o processo de resolução das anáforas é baseado em heurísticas (GOULART; GASPERIN; VIEIRA, 2004). Ela foi desenvolvida em Java e para os seus dados de entrada e saída adotou-se a linguagem de marcação XML, seguindo formatos MMAX (seção 3.3). Utiliza informações adicionais baseadas na análise sintática do analisador sintático PALAVRAS (seção 3.2) para o português.

A arquitetura da ferramenta é baseada em "*pipes & filters*" (GAMMA, 1995), constituindo-se de um conjunto de três etapas com uma ou mais tarefas codificadas através de da linguagem XSL (FIGURA 3.6).

Na primeira etapa, denominada *Input Analysis,* a partir dos arquivos de *Words*, de *POS* e de C*hunks* são extraídos os nodos anafóricos (*anaphor)* (A), onde o atributo *span* corresponde a cada descrição definida do texto (FIGURA 3.7). Nesse passo também são extraídos os nodos candidatos a antecedentes (*candidate)* (B), onde o atributo *span* corresponde aos sintagmas nominais do texto (FIGURA 3.8).

A próxima etapa utiliza a base de resolução de heurísticas (*Resolution Heuristics Base)* composta por um conjunto de regras que implementam as heurísticas para resolução de correferência (regras R_1 à R_n) e filtram os nodos anafóricos (*anaphor)* em busca dos seus antecedentes (VIEIRA; SALMON-ALT; SCHANG, 2002).

Na última etapa, chamada *Output Generation*, os resultados da ferramenta ART são adaptados para o mesmo formato utilizado pela ferramenta MMAX (*elementos markables*) (C) (FIGURA 3.9), com o intuito de possibilitar a comparação dos resultados da marcação automática e da marcação manual (GASPERIN; GOULART; VIEIRA, 2003).

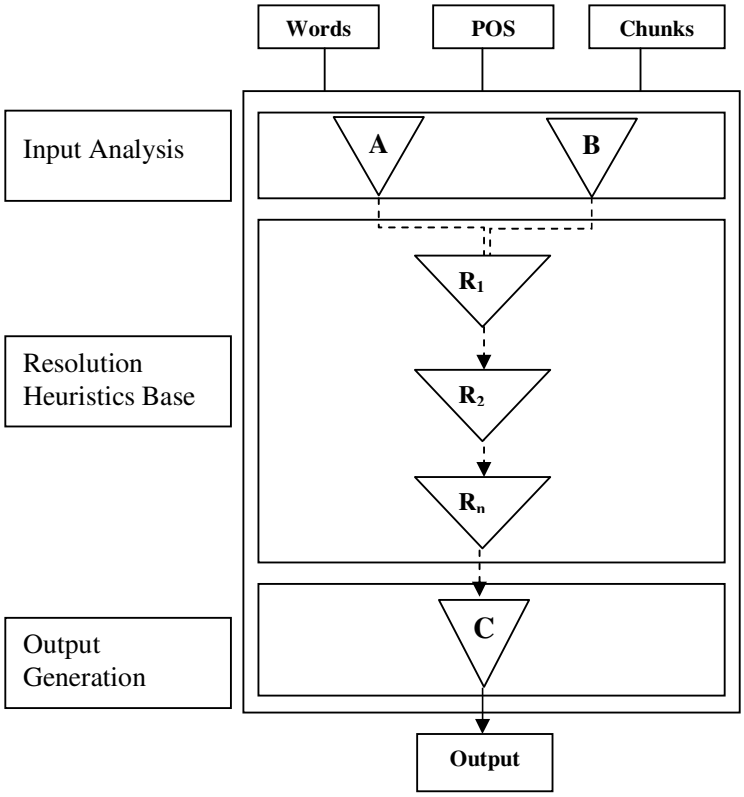

FIGURA 3.6 Arquitetura da Ferramenta ART (GOULART; GASPERIN; VIEIRA, 2004).

Para exemplificar as etapas descritas, retomemos no exemplo (1) a sentença *(ii)* e analisemos as descrições definidas "*o Brasil" (span ="word_708..word_709")* e *"o Othon Palace Hotel" (span ="word_716..word_717)* contidas nesta sentença. Na FIGURA 3.7 é apresentado um trecho do arquivo de anáforas, onde foram extraídas as descrições definidas citadas. Já na FIGURA 3.8 é mostrado um trecho do arquivo de candidatos a antecedentes (todos os sintagmas nominais). Para finalizar, na FIGURA 3.9, é apresentado um trecho do arquivo de marcações, com o resultado da ferramenta ART, correspondente às descrições

definidas tomadas como exemplo, seguindo o mesmo formato da saída do MMAX (FIGURA 3.5).

```
<anaphor-set>
.....
<anaphor span="word_708..word_709"
         pointer="">o Brasil </anaphor>
<anaphor span="word_716..word_717"
         pointer="">o Othon_Palace_Hotel </anaphor>
.....
</anaphor-set>
```

FIGURA 3.7 Arquivo de Anafóricas.

```
<candidate-set>
.....
<candidate span="word_699..word_703">o início de o ano
</candidate>
<candidate span="word_702..word_703">o ano </candidate>
<candidate span="word_708..word_709">o Brasil </candidate>
<candidate span="word_710..word_720">50 discípulos que se
hospedaram em o Othon_Palace_Hotel , em São_Paulo
</candidate>
<candidate span="word_716..word_717">o Othon_Palace_Hotel
</candidate>
.....
</candidate-set>
```

FIGURA 3.8 Arquivo de Candidatos a Antecedentes.

```
<markable>
.....
<markable id="markable_95" span="word_708..word_709"
pointer="markable_36" form="defnp" classification="direct"/>
<markable id="markable_96" span="word_716..word_717"
pointer="" form="defnp" classification="discourse_new"/>
.....
</markable>
```

FIGURA 3.9 Arquivo de Marcações.

3.5 Weka

Nesta seção é descrita a ferramenta Weka que se constitui de um conjunto de implementações de algoritmos de aprendizado de máquina.

O pacote Weka[28] (*Waikato Environment for Knowledge Analysis*) foi desenvolvido pela Universidade de Waikato, na Nova Zelândia e é formado por um conjunto de implementações de algoritmos de Mineração de Dados (WITTEN; FRANK, 2000). O pacote está implementado na linguagem Java, que tem como principal característica ser portável, podendo rodar nas mais variadas plataformas e aproveitar os benefícios de uma linguagem orientada a objetos, como modularidade, polimorfismo, encapsulamento, reutilização de código, dentre outros; além disso, é um software de domínio público.

O Weka possui um formato próprio, o ARFF, que consiste basicamente de duas partes. A primeira parte contém uma lista de todos os atributos, onde se define o tipo do atributo ou os valores que ele pode representar (ao utilizar os valores esses devem estar entre "{}" e separados por vírgulas). A segunda parte consiste das instâncias, ou seja, os registros a serem minerados; com o valor dos atributos para cada instância separado por vírgula; a ausência de um item em um registro deve ser representada pelo símbolo "?". Para exemplificar, a FIGURA 3.10 ilustra um arquivo no formato ARFF.

```
@RELATION descrição_definida
@ATTRIBUTE atributo_1       {TRUE, FALSE}.
@ATTRIBUTE atributo_2       {TRUE, FALSE}
@ATTRIBUTE atributo_3       {TRUE, FALSE}.
@ATTRIBUTE atributo_4       {TRUE, FALSE}.
@ATTRIBUTE class            {nova_discurso, outra}

@DATA
TRUE,FALSE,TRUE,FALSE,nova_discurso
FALSE,TRUE,FALSE,FALSE,nova_discurso
FALSE,FALSE,FALSE,FALSE,outra
FALSE,FALSE,FALSE,FALSE,outra
.................
```

FIGURA 3.10 Arquivo no formato ARFF.

A ferramenta Weka além de possuir métodos de classificação e de predição numérica, dispõe também de 4 tipos diferentes de validação desses métodos:

[28] O Weka está disponível em http://www.cs.waikato.ac.nz/ml/weka/

- *Use trainning set*: não efetuar a divisão. Usar para testar o modelo o mesmo conjunto usado para a sua construção.
- *Supplied test set*: usar um conjunto de teste específico.
- *Cross-Validation n-fold*: validação cruzada[29] em *n-folds*.
- *Percentage split*: treinar o modelo numa percentagem do conjunto de exemplos e testar na restante.

Nesta dissertação, a ferramenta Weka é utilizada no processo de aprendizado, na etapa de classificação das descrições definidas, com a aplicação do algoritmo de árvores de decisão *J48*, que é uma re-implementação em Java do algoritmo C4.5 (QUINLAN, 1993) nessa ferramenta. O algoritmo *j48* não necessita de um ponto inicial de busca; dessa forma pode ser executado apenas uma vez. Entre as funcionalidades do algoritmo *j48* temos: a incorporação de atributos numéricos (*contínuos*); os valores nominais (*discretos*) de um atributo podem ser agrupados de maneira a permitir testes mais complexos; possui poda *(post-pruning)* das árvores para aumentar a precisão; permite lidar com informação incompleta (*missing attribute values*), entre outros. Os valores dos parâmetros de treinamento deste algoritmo podem ser os sugeridos pela ferramenta (fator de confiança de poda (CF) igual a 0,25; número mínimo de instâncias por folha igual a 2 etc.) ou podem ser alterados quando necessário.

Após a escolha do método de classificação e da forma de validação na ferramenta Weka, esta fornece uma avaliação dos resultados com base em algumas medidas de avaliação do desempenho. Algumas destas medidas serão descritas na seção seguinte.

3.6 Avaliação

A avaliação de sistemas de PLN exige um domínio abrangente sobre o problema em observação para possibilitar ou o desenvolvimento de metodologias próprias de avaliação ou

[29] *Validação Cruzada*: os exemplos são aleatoriamente divididos em *r folds* de tamanho *n/r* exemplos. Os exemplos sobre *(r-1) folds* são usados para treinamento e a hipótese induzida é testada no *fold* remanescente.

o bom uso de medidas já existentes. Segundo SANTOS (2001) só é possível o desenvolvimento de uma boa avaliação se o problema analisado for quantificado e suas vantagens de uso identificadas.

Para o desenvolvimento de um método de classificação existem várias questões importantes em relação ao seu desempenho, tanto durante a sua construção como na sua utilização.

No contexto deste trabalho, após o processo de aprendizado (seção 3.9), a resolução das árvores implementadas é avaliada utilizando medidas de desempenho. As medidas de avaliação de desempenho utilizadas aqui são adotadas da área de Recuperação de Informação (RI). Dentre as medidas mais importantes para a avaliação dos resultados e do desempenho estão abrangência, precisão e F-measure (KORFHAGE, 1997; KOWALSKI, 1997); e podem ser obtidas através das seguintes relações:

- Abrangência: é a razão do número de acertos da classe pelo número total de itens da classe, expressa na fórmula mostrada na FIGURA 3.11:

$$\text{Abrangência} = \frac{\text{n\_acertos\_classe}}{\text{n\_classe}}$$

FIGURA 3.11 Abrangência.

- Precisão: representa o grau de confiança de um método de classificação, o que torna a medida mais importante e necessária em qualquer tarefa de indução. É geralmente calculada como a razão entre o número de exemplos corretamente classificados sobre o número total de exemplos classificados na classe, ilustrada na fórmula descrita na FIGURA 3.12:

$$\text{Precisão} = \frac{\text{n\_acertos\_classe}}{\text{n\_classificados\_classe}}$$

FIGURA 3.12 Precisão.

A análise de avaliação deve considerar tanto precisão quanto abrangência. Por isso, uma medida que combine os valores de precisão e de abrangência obtendo o desempenho geral do sistema, deve ser adotada. Neste trabalho, empregamos a medida F-measure que permite um balanceamento entre os valores de precisão e abrangência através da expressão na seguinte fórmula (FIGURA 3.13) em que β é o parâmetro que permite a atribuição de diferentes pesos para as medidas de precisão (P) e abrangência (C), sendo 1 o valor freqüentemente utilizado. O valor do F-measure (F) é maximizado quando a precisão e a abrangência são iguais ou muito próximas, assumindo por definição, o próprio valor da precisão ou da abrangência, representando o ponto de equilíbrio do sistema.

$$F = \frac{(\beta^2 + 1) * P*C}{\beta^2 * (P + C)}$$

FIGURA 3.13 F-measure.

Além dessas medidas, utilizamos o percentual de acertos para a avaliação.

3.7 Grupos de Características das descrições definidas

Um estudo de corpus foi realizado para analisar aspectos estruturais das descrições definidas com o objetivo de identificar as características que podem sinalizar as expressões novas no discurso. Este estudo partiu de trabalhos realizados na Língua Inglesa (VIEIRA, 1998), de um estudo inicial para a Língua Portuguesa (COLLOVINI; GOULART; VIEIRA, 2004) e da análise das possíveis configurações para as descrições definidas da Língua Portuguesa (seção 2.4). Dessa forma, algumas características previamente identificadas para o inglês foram verificadas para a Língua Portuguesa. Além destas, novas características foram adicionadas.

As características foram organizadas em três grupos para examinar melhor o desempenho de cada aspecto. O primeiro grupo é composto por características relacionadas especificamente à estrutura da descrição definida. No segundo grupo, reunimos as características que envolvem a análise da sentença. E o terceiro baseia-se na análise do texto.

Dessa forma, com a análise de corpus observamos que a primeira manifestação de uma descrição definida ocorre de modo mais completo do que as demais. Por isso, os elementos complementares – elementos que se integram ao nome núcleo para completá-lo ou aperfeiçoá-lo – foram selecionados como características do primeiro grupo (Grupo 1).

Analisamos também as relações que ocorrem em descrições definidas com verbos de ligação (ser, estar, permanecer etc.) denominadas *construções copulares*, pois composições mais completas podem sinalizar a primeira manifestação de um referente no texto. A verificação de construções copulares compõe o segundo grupo de características (Grupo 2). Entendendo que expressões da primeira sentença não possuem antecedentes, elaboramos uma característica que analisa a posição da sentença e que faz parte do segundo grupo (Grupo 2). Seguindo o princípio de que descrições definidas novas no discurso não apresentam antecedentes foi desenvolvida, então, uma característica que busca um antecedente no texto (Grupo 3).

Estes grupos de características das descrições definidas são descritas a seguir.

1. **Características baseadas na estrutura do sintagma (GRUPO 1):**

• Característica 1 (SP): descrições definidas acompanhadas de um sintagma preposicional, pós-modificador (restritivo). Por exemplo:

"*A tarde <u>de</u> <u>ontem</u>*".

Nesses casos, um pós-modificador restritivo sucede o núcleo restringindo-o. Um modificador restritivo permite que o referente seja identificado através da informação do modificador (neste caso, um sintagma preposicional) que especifica a informação do núcleo, não sendo necessário recorrer a um antecedente para a sua interpretação.

• Característica 2 (APO): descrições definidas constituídas de construções de apostos com marca explícita. Por exemplo:

"*O prefeito de Gravataí, Daniel Luiz Bordignom*".

No corpus estudado, por tratar-se de textos jornalísticos, são relatadas informações explicativas (sobre locais, pessoas, empresas, eventos etc.), sendo que uma característica observada nesses textos é a presença de construções de aposto que geralmente introduzem um novo referente no discurso, sendo auto-explicativas.

• Característica 3 (APO_NP): descrição definida acompanhada de um nome próprio constituindo um aposto sem marca explícita. Por exemplo:

"*O delegado Elson Campelo*".

No corpus estudado, uma característica observada nos textos é a construção de descrições definidas com núcleo sendo um nome comum (substantivo comum), seguido de um nome próprio com função de aposto e geralmente tratando-se de um novo referente no discurso.

• Característica 4 (REL): descrições definidas acompanhadas de uma cláusula relativa. Por exemplo:

"*O texto que deve ser assinado pelos jornalistas*".

No corpus estudado uma característica observada nos textos é a presença de cláusulas relativas modificando um nome e que geralmente introduzem um novo referente no discurso.

• Característica 5 (NP_COM): descrição definida com o núcleo sendo um nome próprio composto. Por exemplo:

"O Othon Palace Hotel".

No corpus estudado, verificamos que as estruturas com nomes próprios compostos coincidem, predominantemente, com a classe nova no discurso. Geralmente, na continuidade do texto, as expressões que retomam descrições definidas com nomes próprios compostos o fazem com apenas um dos nomes próprios ou com termos sinônimos.

• Característica 6 (SA): descrição definida acompanhada de um sintagma adjetival, pós-modificador (restritivo). Por exemplo:

"*As conversas mais antigas*".

Um pós-modificador restritivo sucede o núcleo limitando-o e permite que o referente seja identificado através da informação do modificador (neste caso, um sintagma adjetival) que especifica a informação do núcleo.

• Característica 7 (PRE_ADJ): descrição definida que apresenta um adjetivo anteposto ao núcleo, pré-modificador (restritivo). Por exemplo:

"*O primeiro grau*".

Um pré-modificador restritivo antecede o núcleo, limitando-o e especificando a informação do núcleo.

• Característica 8 (PRE_NUM): descrição definida composta por um numeral anteposto ao núcleo, pré-modificador (restritivo). Por exemplo:

"*Os 65 anos*".

Essa característica entende como numeral os números cardinais, ordinais, multiplicativos e fracionários.

• Característica 9 (NUM): descrição definida que apresenta após o núcleo um numeral, pós-modificador (restritivo). Por exemplo:

"*Os anos 60*".

Os números cardinais, ordinais, multiplicativos e fracionários fazem parte do conceito de numeral dessa característica.

• Característica 10 (DET): descrição definida que possui, além do artigo definido, outro(s) determinante(s). Por exemplo:

"*Os nossos arqueólogos*".

Ao lado do artigo definido, podem atuar como determinantes pronomes indefinidos (todos, outros, poucos etc.), possessivos (seu, nossos, meu etc.), demonstrativos (mesmo, outro, demais etc.).

• Característica 11 (SUP): descrição definida acompanhada de um superlativo. Por exemplo:

"*Os melhores alunos*".

Superlativo é uma palavra que expressa uma qualidade (adjetivo) em um grau elevado ou no mais elevado grau.

• Característica 12 (SUP_A): descrição definida que descreve o grau máximo de qualidade (adjetivo), representando o maior índice de uma escala. Por exemplo:

"*O Christofle líquido é o melhor*".

Uma descrição definida com essa característica desempenha a função de predicativo do sujeito – que é o termo ou expressão que complementa o sujeito. Para essa característica, a formação do predicativo do sujeito é constituída por verbo de ligação (ser, estar, permanecer etc) mais uma descrição definida composta por não mais que dois termos (ARTIGO DEFINIDO + ADJETIVO).

• Característica 13 (TAM): descrição definida composta por cinco ou mais termos. Por exemplo:

"*As mais recentes criações estéticas brasileiras*".

2. **Características baseadas na análise da sentença (contexto sintático e posição da sentença) (GRUPO 2):**

• Característica 14 (COP): descrições definidas em uma construção copular. Por exemplo:

"*O coreano seria a língua dos anjos*".

Nesses casos, uma construção copular é uma relação entre dois sintagmas nominais por meio de um verbo de ligação (ser, estar, permanecer etc.).

- Característica 15 (PRI_SENT): descrições definidas que ocorrem na primeira sentença do texto.

A primeira sentença do texto pode ser o título ou, na ausência deste, a primeira frase do primeiro parágrafo do texto.

3. **Características baseadas na análise do texto (GRUPO 3):**

- Característica 16 (SEM_ANT): o núcleo da descrição definida é uma palavra que não ocorre anteriormente no texto.

De posse das características das descrições definidas, apresentaremos a seguir, a forma de identificação automática destas nos corpora estudado.

3.8 Identificação Automática das Características no Corpus

Para a identificação automática das características apresentadas na seção 3.7 são necessárias informações da estrutura sintática das descrições definidas, presentes no arquivo de *chunks* e geradas pela ferramenta Xtractor (seção 3.2), onde a quantidade de informação sintática de interesse pode variar de acordo com as características utilizadas.

Primeiramente, apresentaremos como são representadas as informações de interesse das descrições definidas (do arquivo de *chunks*) e depois discutiremos a extração dessas informações com a identificação automática das características usando a linguagem XSL[30] (*eXtensible Stylesheet Language*). Para exemplificar, então retomamos algumas das características (1, 5, 6) apresentadas na seção 3.7.

[30] Desenvolvida pelo W3C (*World Wide Web Consortium*) disponiível em: http://www.w3.org/Style/XSL/

Na característica 1, procura-se a existência de um sintagma preposicional no *chunk* da descrição definida, ou seja, um filho desse *chunk* com o atributo *form* igual a "*pp*". A FIGURA 3.14 ilustra o *span* "*word_200..word_203*" que corresponde a *"a tarde de ontem"*.

```
........
<chunk id="chunk_277" ext="p" form="np"
        span="word_200..word_203">
  <chunk id="chunk_278" ext="n" form="art" span="="word_200">
  </chunk>
  <chunk id="chunk_279" ext="h" form="n" span="="word_201">
  </chunk>
  <chunk id="chunk_280" ext="n" form="pp"
          span="word_202..word_203">
     <chunk id="chunk_281" ext="h" form="prp" span="word_202">
     </chunk>
     <chunk id="chunk_282" ext="p" form="adv" span="word_203">
     </chunk>
  </chunk>
........
```

FIGURA 3.14 Trecho do Arquivo de *Chunks.*

Na característica 5, procura-se o núcleo dessa estrutura, ou seja, o filho desse *chunk* que possua o atributo *ext* igual a "*h*" e a forma de nome próprio, isto é, o atributo *form* igual a "*prop*". A FIGURA 3.15 ilustra o *span "word_72..word_73"* correspondente a *"a Rádio_Globo_Washington_Rodrigues"*.

```
.....
<chunk id="chunk_100" ext="p" form="np"
       span="word_72..word_73">
  <chunk id="chunk_102" ext="n" form="art" span="word_72">
  </chunk>
  <chunk id="chunk_102" ext="h" form="prop" span="word_73">
  </chunk>
.....
```

FIGURA 3.15 Trecho do Arquivo de *Chunks.*

Já a característica 6 verifica a presença de um sintagma adjetival no *chunk* da descrição definida, ou seja, o filho desse *chunk* que possua o atributo *form* igual a "*ap*". A FIGURA 3.16 ilustra o *span* "*word_22..word_28*" que corresponde a *"os momentos mais difíceis de minha carreira"*.

```
.....
<chunk id= “chunk_31" ext="p" form="np"
       span= “word_22..word_28”>
  <chunk id="chunk_32" ext="n" form="art" span="word_22">
  </chunk>
  <chunk id="chunk_33" ext="h" form="n" span="word_23">
  </chunk>
  <chunk id="chunk_34" ext="n" form="ap"
         span=“word_24..word_25”>
    <chunk id="chunk_35" ext="a" form="adv" span="word_24">
    </chunk>
    <chunk id="chunk_36" ext="h" form="adj" span="word_25">
    </chunk>
</chunk>
.....
```

FIGURA 3.16 Trecho do Arquivo de *Chunks.*

Para a extração das informações da estrutura sintática do arquivo de *chunks* foi utilizada a linguagem XSL. XSL é um conjunto de instruções destinadas à visualização de documentos XML. Desta forma, a partir de um arquivo XSL é possível transformar um documento XML em diversos formatos, incluindo RTF, TeX, PostScript, HTML e TXT. A linguagem XSL auxilia a identificação dos elementos (*nodos*) de um documento XML, permitindo a simplificação do processamento de transformação desses elementos em outros formatos de apresentação. Contudo, é possível criar múltiplas representações da mesma informação a partir de vários documentos XSL aplicados a um único documento XML.

Para mostrar como se dá a extração destas informações de interesse do arquivo de *chunks* são apresentadas no **APÊNDICE A** as folhas de estilos, correspondente a implementação da característica 1, da característica 5 e da característica 6.

De posse do corpus anotado e das características levantadas, podemos realizar experimentos em uma ferramenta de resolução de anáforas, como a ferramenta ART, ou gerar as entradas para os experimentos com AM.

Na seção seguinte será relatado o processo de aprendizado utilizando a extração automática das características das descrições definidas apresentada.

3.9 Processo de Aprendizado

Nesta dissertação, algoritmos de AM Supervisionado serão utilizados para a indução de árvores de decisão para a classificação de descrições definidas. Para isso, é proposta uma metodologia similar às etapas tradicionais de Sistema de Categorização de Textos (SCT), ilustrada na FIGURA 3.17 e descrita a seguir.

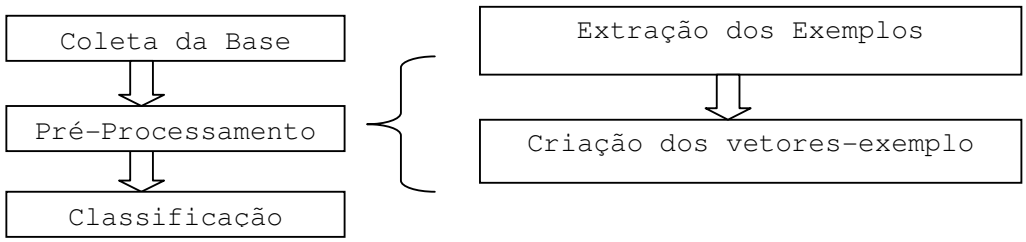

FIGURA 3.17 Metodologia Proposta.

Nas próximas seções a metodologia proposta para o processo de aprendizagem no contexto deste estudo é descrita.

3.9.1 Coleta da Base

A coleta da base consiste na obtenção dos exemplos a serem utilizados para o treinamento do classificador. O exemplo descreve o objeto de interesse e é formado por um vetor de valores dos atributos. A base de dados usada no treinamento do classificador é formada por um conjunto de 24 textos da Folha de São Paulo (corpus 1) e os exemplos utilizados pelo classificador são as descrições definidas presentes nesses textos (seção 3.1).

3.9.2 Pré-processamento

A etapa de pré-processamento é a mais custosa por possuir várias operações sobre os textos, objetivando a criação de uma representação computacional da base de dados seguindo

algum modelo. Segundo a metodologia proposta neste trabalho, o pré-processamento é formado por duas etapas: extração dos exemplos e criação dos vetores-exemplo.

A primeira etapa do pré-processamento, a extração das descrições definidas da base de dados, tem por objetivo selecionar apenas as descrições definidas identificadas tanto na marcação manual e na automática[31]. Para isso, podemos dividir o processo de extração em dois passos:

1. A partir do corpus anotado manualmente, é verificado quais elementos *markables* do arquivo de marcação manual apresentam atributo *form* igual a *"defnp"*, ou seja, é uma descrição definida;

2. De posse das descrições definidas (passo 1) é verificado se cada descrição definida selecionada apresenta os mesmos elementos (atributo *span*) na marcação manual (seção 3.3) e na marcação automática (seção 3.2).

Para exemplificar o processo de extração dos exemplos, é ilustrado na FIGURA 3.18 um trecho do arquivo de marcação manual (*markables*) e na FIGURA 3.19 um trecho do arquivo de marcação automática (*chunks*) referente à descrição definida: *"o bicheiro Castor de Andrade"*.

Observamos primeiramente que a expressão analisada possui a forma de artigo definido *(form = "defnp")* no arquivo de marcação (FIGURA 3.18), representando assim que se trata de uma descrição definida. Após, verificamos que a expressão possui o mesmo atributo *span ("word_128..word_130")* no arquivo de *chunks* (FIGURA 3.19) e no arquivo de marcações, representando que se tratam, em ambos os arquivos, da mesma expressão.

[31] Idealmente, a extração das descrições definidas deveria ser feita automaticamente com base no corpus anotado. Isso foi feito apenas para o corpus 2. No corpus 1 a extração das descrições definidas foi manual para a marcação manual e automática para a marcação automática. Porém ocorreram diferenças entre os dois conjuntos de descrições definidas, então, uma verificação posterior foi necessária.

```
...
<markable id="markable_19" span="word_128..word_130"
          coreference="non_coreferent" form="defnp"
          classification="discourse_new" />
...
```

FIGURA 3.18 Arquivo de marcações.

```
...
<chunk id="chunk_181" ext="p" form="np"
      span="word_128..word_130">
  <chunk id="chunk_182" ext="n" form="art" span="word_128">
  </chunk>
  <chunk id="chunk_183" ext="h" form="n" span="word_129">
  </chunk>
  <chunk id="chunk_184" ext="n" form="prop" span="word_130">
  </chunk>
</chunk>
...
```

FIGURA 3.19 Arquivo de *chunks.*

Para a verificação e extração dessas informações de interesse (descritas no passo 1 e no passo 2) foi construída uma folha de estilos XSL, a qual é executada para cada texto da base de dados, tendo como resultado uma lista das descrições definidas do referente texto. Para ilustrar, a FIGURA 3.20 apresenta um trecho de um texto do corpus utilizado e na FIGURA 3.21 parte da lista de descrições definidas resultante.

```
...
Citados negam as acusações. Da Sucursal do Rio. O ex-deputado e
cantor Agnaldo Timóteo afirmou que a presença de seu nome na
lista do banqueiro do jogo do bicho Castor de Andrade mostra o
quanto o contraventor é organizado. Sempre que estou duro, peço
dinheiro a ele. Castor é meu pai branco.
...
```

FIGURA 3.20 Trecho de um texto.

```
as acusações
Da_Sucursal_do_Rio
o jogo_do_bicho
o contraventor
...
```

FIGURA 3.21 Lista com algumas descrições definidas.

Como resultado da primeira etapa do pré-processamento, temos uma base de dados para os experimentos, constituída de 1105 descrições definidas do corpus 1. Desta forma para os experimentos realizados, foi utilizada parte do conjunto de exemplos do corpus 1; das 1411 descrições definidas foram utilizadas as 1105 resultantes da primeira etapa do pré-processamento.

Para a segunda etapa do processo de pré-processamento, temos a criação dos vetores-exemplo com base no conjunto de exemplos (descrições definidas) da base de dados extraídos na etapa anterior. Um conjunto de exemplos é composto por vetores-exemplo contendo os valores dos atributos e a classe associada (MONARD, BARANAUSKAS, 2003). No contexto deste trabalho, os atributos correspondem às características das descrições definidas (seção 3.7) e como se objetiva identificar as descrições definidas novas no discurso, a classe utilizada para cada descrição definida será a classificação "*nova_ discurso*" ou "*outra*".

Para a criação dos vetores-exemplo é necessário definir um tipo de codificação e representá-lo no modelo de espaço vetorial. O tipo de codificação utilizado é a codificação "*binária*", onde cada posição do vetor-exemplo indica a presença (*TRUE*) ou a ausência (*FALSE*) de determinado atributo no exemplo.

Para exemplificar o processo de criação dos vetores-exemplo, retomamos a descrição definida "*o bicheiro Castor de Andrade*". A partir dessa expressão, primeiramente é checado o valor dos seus atributos utilizando o arquivo de *Chunks* (FIGURA 3.19) e assim verifica-se a presença/ausência de cada uma das características das descrições definidas (seção 3.7) utilizando o processo descrito anteriormente na seção 3.8. Como resultado, o vetor-exemplo correspondente é gerado.

Após a verificação dos atributos, é necessária a informação da classe correspondente ao exemplo analisado. Esta informação é possível com o auxílio do corpus anotado (arquivo de *markables*), onde será analisado o atributo "*classification*" que corresponde à classificação

anafórica da expressão. Como resultado, temos para a descrição definida "*o bicheiro Castor de Andrade*" o seguinte vetor-exemplo: *TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, nova_discurso.*

3.9.3 Classificação

Após a etapa de pré-processamento, é iniciada a etapa de classificação com árvores de decisão, utilizando a ferramenta Weka. Essa ferramenta re-implementa o algoritmo C4.5 denominando-o de *j48* (seção 3.5).

Na aplicação do algoritmo *j48* neste trabalho, é realizada a simplificação das árvores de decisão geradas (seção 2.5). O parâmetro chave para a simplificação das árvores geradas é o nível de confidência de poda, também chamado de fator de confiança de poda (CF), o qual pode sofrer ajustes (valor padrão de CF é 0,25 ou 25%) com o objetivo de evitar o *overfitting* dos dados de treinamento. Esse ajuste depende da quantidade e da confiabilidade dos dados de treinamento disponíveis. Nesse trabalho, foram testados diferentes valores para o fator de confiança de poda (CF), resultando na utilização dos valores: CF = 0,10 e CF = 0,35. Os demais valores dos parâmetros do algoritmo *j48* foram os sugeridos pela ferramenta. Cabe ressaltar que, dos métodos apresentados, o usado foi validação cruzada (*cross-validation*) em *10 folds* na base de dados (seção 3.9.2).

Para a execução do algoritmo *j48* na ferramenta Weka é necessário arquivos de entrada respeitando o formato ARFF (seção 3.5). Estes arquivos de entrada são formados pelos vetores-exemplo gerados na etapa de pré-processamento (seção 3.9.2), respeitando o tipo de codificação e de classificação utilizados. Para exemplificar, a FIGURA 3.22 mostra um trecho de um arquivo de entrada para a ferramenta Weka. No **APÊNDICE A** é

apresentada a folha de estilos XSL que gera os arquivos de entrada para a ferramenta Weka no formato ARFF.

As categorias ("*nova_discurso*" ou "*outra*") corretamente e incorretamente classificadas são mostradas em porcentagem, as medidas de precisão, abrangência e F-measure são apresentadas nos detalhes da acurácia por categoria (seção 3.6). Também é mostrada a matriz de confusão com a quantidade de descrições definidas classificadas por categoria ("*nova no discurso*" ou "*outra*").

```
@relation entrada
@ attribute SEM_ANT {TRUE,FALSE}
@attribute PRE_ADJ {TRUE,FALSE}
@attribute PRE_NUM {TRUE,FALSE}
@attribute NUM {TRUE,FALSE}
@attribute REL {TRUE,FALSE}
@attribute NP_COM {TRUE,FALSE}
@attribute COP {TRUE,FALSE}
@attribute SP {TRUE,FALSE}
@attribute SA {TRUE,FALSE}
@attribute APO {TRUE,FALSE}
@attribute APO_NP {TRUE,FALSE}
@attribute SUP {TRUE,FALSE}
@attribute SUP_A {TRUE,FALSE}
@attribute PRI_SENT {TRUE,FALSE}
@attribute TAM {TRUE,FALSE}
@attribute DET {TRUE,FALSE}
@attribute category {nova_discurso,outra}
@data
FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,TRUE,FALSE,FALSE,FALSE,
FALSE,FALSE,FALSE,TRUE,FALSE,FALSE,outra
TRUE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,
FALSE,FALSE,FALSE,TRUE,FALSE,FALSE,nova_discurso
.....................
```

FIGURA 3.22 Trecho do Arquivo de Entrada no Formato ARFF.

Uma vez apresentado o processo de aprendizado, passemos, então, aos passos necessários para a implementação da classificação automática no ambiente ART.

3.10 Implementação no Ambiente ART

Após o processo de aprendizado (seção 3.9), os experimentos foram executados com a ferramenta Weka (aplicação do algoritmo *j48*) e geradas as árvores de decisão. A

classificação automática no ambiente ART baseou-se nos atributos (características das descrições definidas) presentes nessas árvores. Além disso, uma posterior avaliação da classificação automática no ambiente ART foi desenvolvida com base na classificação manual presente no arquivo de marcações (*markables)*.

Nesta seção são apresentados os passos seguidos para a execução da classificação automática das descrições definidas no ambiente ART. Os experimentos realizados com árvores de decisão; a combinação dos atributos utilizados e a análise da classificação, utilizando as medidas de avaliação (seção 3.6), são mostradas no capítulo 4.

Para a implementação da classificação automática das descrições definidas no Ambiente ART, foi utilizada a arquitetura da ferramenta, descrita na seção 3.4, baseada em *"pipes&filters"*, seguindo uma seqüência de quatro etapas:

1. A entrada para a ferramenta é selecionada, que constitui do conjunto de exemplos da base de dados (extraídos do corpus 1 e corpus 2);

2. Seleção e aplicação de um conjunto de regras, sendo que as regras são a implementação das características das descrições definidas (seção 3.8);

3. O resultado da aplicação das regras da etapa anterior é representado para o formato de saída da ferramenta MMAX (seção 3.3);

4. Comparação dos resultados da classificação automática no ambiente ART (atributo "*classification*") e da classificação manual na ferramenta MMAX (atributo "*classification*") a partir da aplicação de folhas de estilos XSL. O resultado desta avaliação é apresentado na seção 4.2.

3.11 Considerações Finais

Neste capítulo foi apresentada a descrição do corpus e das ferramentas utilizadas no trabalho (PALAVRAS, Xtractor, MMAX, Weka, ART).

Os grupos de características das descrições definidas originados do estudo de corpus foram mostrados na seção 3.7. As características apresentadas em estudos anteriores da Língua Inglesa foram analisadas e algumas foram aproveitadas. Por exemplo, a presença do nome próprio (*Proper names*) em uma descrição definida era uma característica para a classificação desta como nova no discurso em trabalhos relacionados. Neste trabalho, a característica *"nome próprio"* sofreu alterações. Uma delas foi analisar se o núcleo da descrição definida é um nome próprio composto (veja Característica 5: NP_COM) para tratar casos como "*O Ministério das Relações Exteriores*" (nova no discurso) e "*O Ministério*" (outra). Outra alteração resultou na característica "*aposto nome próprio"* (veja Característica 3: APO_NP) que combina duas características e analisa se a descrição definida é formada por um nome próprio que desempenha a função de aposto.

Entre as características das descrições definidas aproveitadas, temos as características: *"sintagma preposicional"* (veja Característica 1: SP), "*construções de apostos"* (veja Característica 2: APO), *"cláusula relativa"* (veja Característica 4: REL), *"sintagma adjetival"* (veja Característica 6: SA); *"pré modificador adjetivo"* (veja Característica 7: PRE_ADJ), *"pré modificador número"* (veja Característica 8: PRE_NUM), *"pós modificador número"* (veja Característica 9: NUM), *"superlativo"* (veja as Características 11: SUP), "*construções copulares"* (veja Característica 14: COP), *"primeira sentença"* (veja Característica 15: PRI_SENT) e *"sem antecedente"* (veja Característica 16: SEM_ANT).

Outras características novas foram identificadas. Por exemplo, as características que analisam a presença de um outro determinante (veja Característica 10: DET) e de um superlativo absoluto (veja Característica12: SUP_A). Destaca-se também, a característica *"tamanho"* (veja Característica 13: TAM) por ser uma característica original não verificada em nenhum outro estudo.

Nesta dissertação, no processo de aprendizado, para a geração das árvores de decisão, adotou-se uma metodologia similar às etapas tradicionais de Sistemas de Categorização de Textos. Esta metodologia tem como resultado um conjunto de vetores-exemplo que são as entradas para a ferramenta Weka. Após o processo de aprendizado com o corpus da Folha de São Paulo (corpus 1), as árvores de decisão geradas são avaliadas utilizando medidas de desempenho (seção 3.6). A classificação automática das descrições definidas no ambiente ART utilizará os atributos considerados relevantes nas árvores de decisão analisadas.

4. Resultados

Neste capítulo são apresentados os resultados deste estudo. Primeiramente foi realizado o processo de aprendizado (seção 3.9) com a construção da base de dados e a classificação utilizando árvores de decisão. Os atributos e os parâmetros utilizados para a geração das árvores de decisão na ferramenta Weka são descritos na seção 4.1, juntamente com a descrição dos experimentos realizados e respectivos resultados. Além disso, neste capítulo, é apresentada, na seção 4.2, a avaliação das árvores aprendidas no ambiente ART.

Os experimentos foram divididos em:

- Experimento 1: classificação das descrições definidas como novas no discurso ou outra;
- Experimento 2: classificação das descrições definidas como não correferentes (englobando a classe nova no discurso e anafórica associativa) ou outra (englobando as demais classificações).

4.1 Geração de Árvores de Decisão

Nos experimentos foram utilizados como atributos as características das descrições definidas, de acordo com três grupos (G1, G12 e G123) apresentados a seguir:

- G1: características do GRUPO 1 (estrutura do sintagma nominal), totalizando 13 atributos;
- G12: características do GRUPO 1 (estrutura do sintagma nominal) e do GRUPO 2 (análise da sentença), totalizando 15 atributos;

- G123: características do GRUPO 1 (estrutura do sintagma nominal), do GRUPO 2 (análise da sentença) e do GRUPO 3 (análise do texto), totalizando 16 atributos.

Nos experimentos foram utilizados como fator de confiança de poda (CF) os valores: CF = 0,10 e CF = 0,35. Com CF = 0,10 a árvore resultante é simplificada e os atributos que não contribuíram significativamente para a classificação são retirados. Já com a variação para CF = 0,35 ocorre um aumento do número de atributos presentes nas árvores de decisão, sendo possível uma análise mais completa dos atributos utilizados para a classificação.

Nas próximas seções são apresentados os resultados dos experimentos realizados na Ferramenta Weka (as árvores de decisão geradas estão no **APÊNDICE B).**

4.1.1 Experimento 1 – novas no discurso (Weka)

Para analisar os resultados, são utilizadas as medidas de avaliação e de desempenho, descritas na seção 3.6. As informações referentes à base de dados utilizada (corpus 1) são apresentadas na Tab. 4.1.

TABELA 4.1 Número de exemplos de cada classe.

Classes adotadas	Classes	Nº Exemplos	Totais
nova_discurso	nova_discurso	550	**550**
outra	direta	285	**555**
	indireta	159	
	associativa	94	
	outra	17	
Total de exemplos			**1105**

Para avaliação dos resultados deste experimento, é utilizado como *baseline* um algoritmo que considera todas as descrições definidas como novas no discurso. Este *baseline* é apresentado na Tab 4.2.

TABELA 4.2 Resultados do *baseline.*

Atributos	% Acertos	Classes	Precisão	Abrangência	F-measure
baseline	49,7%	nova_discurso	0,50	1,00	0,66
		outra	0	0	0

Atributos G1:

Os resultados com os atributos G1 são apresentados na Tab. 4.3.

TABELA 4.3 Experimento 1 com atributos G1.

CF	Acertos	Classe	Precisão	Abrangência	F-Measure
0,10	63,0%	nova_discurso	**0,66**	0,54	0,59
		outra	0,61	0,72	0,66
0,35	63,0%	nova_discurso	0,65	**0,55**	**0,60**
		outra	0,61	0,70	0,65

Obtivemos para a classe nova no discurso uma precisão em torno de 66%, uma abrangência em torno de 55%, resultando uma F-measure em torno de 60%. Não há uma grande diferença nos resultados em relação ao fator de confiança de poda. A classe outra não foi analisada neste estudo.

Nas duas árvores de decisão geradas, são comuns os atributos: TAM, SA, NP_COM, PRE_ADJ. A segunda árvore (CF = 0,35), além desses atributos, apresenta também o atributo APO_NP.

Atributos G12:

Os resultados obtidos com os atributos G12 são mostrados na Tab. 4.4.

TABELA 4.4 Resultados do Experimento 1 com atributos G12.

CF	Acertos	Classe	Precisão	Abrangência	F-Measure
0,10	63,8%	nova_discurso	0,66	0,55	0,60
		outra	0,62	0,73	0,67
0,35	63,8%	nova_discurso	**0,66**	**0,57**	**0,61**
		outra	0,62	0,71	0,66

Para a classe nova no discurso alcançamos 66% em precisão e abrangência em torno de 57%, resultando uma F-measure em torno de 61%.

Nas duas árvores de decisão resultantes, estão presentes os atributos: PRI_SENT, TAM, SA, PRE_ADJ, NP_COM. A árvore com CF = 0,35 possui também o atributo APO_NP.

O atributo PRI_SENT foi incluído no topo da árvore, já o atributo COP não foi considerado. Não há uma grande diferença nos resultados dos grupos G1 e G12.

Atributos G123:

Os resultados obtidos com os atributos G123 são ilustrados na Tab. 4.5.

TABELA 4.5 Resultados do Experimento 1 com G123 de atributos.

CF	Acertos	Classe	Precisão	Abrangência	F-Measure
0,10	70,4%	nova_discurso	**0,65**	**0,88**	**0,75**
		outra	0,82	0,53	0,64
0,35	69,1%	nova_discurso	0,64	0,87	0,74
		outra	0,80	0,52	0,63

Observamos para a classe nova no discurso uma precisão em torno de 65%, similar aos grupos anteriores. Pode-se observar um ganho nos valores de abrangência (88%) e F-measure (75%).

Nas duas árvores de decisão geradas, os atributos comuns são: PRI_SENT, SEM_ANTE, SUP, NUM. A segunda árvore (CF = 0,35), além desses atributos, apresenta também os atributos SA, COP, SP TAM. Aqui podemos observar que os dois atributos novos de G12 foram incluídos (PRI_SENT e COP).

Na Tab. 4.6 é apresentada a taxa de acertos e F-measure para cada combinação de atributos utilizada, destacando-se os resultados dos atributos G123.

TABELA 4.6 Taxa de acertos do Experimento 1.

Atributos	CF	% acertos	F-measure
G1	0,10	63,0%	0,59
	0,35	63,0%	0,60
G12	0,10	63,8%	0,60
	0,35	63,8%	0,61
G123	0,10	**70,4%**	**0,75**
	0,35	**69,1%**	**0,74**

Comparamos os melhores resultados do Experimento 1 com o *baseline* na Tab. 4.7.

TABELA 4.7 Comparação entre o *baseline* e o Experimento 1

Atributos	% Acertos	Classes	Precisão	Abrangência	F-measure
BASELINE	49,7%	nova_discurso	**0,50**	**1,00**	**0,66**
		outra	0	0	0
G1	63,0%	nova_discurso	0,65	0,55	0,60
		outra	0,61	0,70	0,65
G12	63,8%	nova_discurso	**0,66**	0,57	0,61
		outra	0,62	0,71	0,60
G123	70,4%	nova_discurso	0,65	**0,88**	**0,75**
		outra	0,82	0,53	0,64

Com 66% de precisão e 75% de F-measure obtivemos para a classe nova no discurso um ganho em relação ao *baseline* (50%, 66%). Além disso, verificaram-se ganhos significativos na taxa de acertos (de 49,7% para 70,4%).

2.1.2 Experimento 2 – não correferentes (Weka)

Os resultados do Experimento 2 foram avaliados de forma similar ao Experimento 1. Informações referentes à base de dados utilizada são apresentadas na Tab. 4.8.

TABELA 4.8 Número de exemplos de cada classe.

Classes adotadas	Nº Exemplos
não_correferente	661
outra	444
Total	**1105**

Para a avaliação dos resultados deste experimento, é utilizado como *baseline* um algoritmo que considera todas as descrições definidas como não correferentes (Tab 4.9).

TABELA 4.9 Resultados do *baseline.*

Atributos	% Acertos	Classes	Precisão	Abrangência	F-measure
baseline	59,8%	não_correferente	0,60	1,00	0,75
		outra	0	0	0

Atributos G1:

Os resultados obtidos com os atributos G1 são apresentados a seguir na Tab. 4.10.

TABELA 4.10 Resultados do Experimento 2 com atributos G1.

CF	% Acertos	Classe	Precisão	Abrangência	F-Measure
0,10	61,3%	não_correferente	**0,71**	**0,58**	**0,64**
		outra	0,53	0,66	0,59
0,35	61,4%	não_correferente	0,70	0,58	0,64
		outra	0,53	0,66	0,59

Obtivemos para a classe não correferente precisão em torno de 71%, abrangência de 58%, resultando uma F-measure de 64%.

Nas duas árvores de decisão geradas, os atributos comuns são: TAM, NUM, APO_NP, SA, NP_COM, PRE_ADJ, SP. A segunda árvore (CF = 0,35), além desses atributos, apresenta também o atributo PRE_NUM.

O conjunto de atributos difere daqueles considerados para o Experimento 1 (novas no discurso).

Atributos G12:

Os resultados com os atributos G12, são apresentados a seguir na Tab. 4.11.

TABELA 4.11 Resultados do Experimento 2 com atributos G12.

CF	% Acertos	Classe	Precisão	Abrangência	F-Measure
0,10	63,0%	não_correferente	**0,71**	**0,61**	**0,66**
		outra	0,55	0,66	0,60
0,35	62,4%	não_correferente	0,70	0,61	0,65
		outra	0,54	0,64	0,59

Para a classe não correferente foi alcançada uma precisão em torno de 71%, abrangência de 61%, resultando F-measure em torno de 66%. Aqui se pode observar um pequeno ganho em relação a G1.

Nas duas árvores de decisão geradas, os atributos comuns são: PRI_SENT, TAM, APO_NP, NUM, SA, NP_COM, PRE_ADJ, SP. A árvore considerando CF = 0,35, além desses atributos, apresenta também o atributo COP. Podemos observar que os atributos G2 (PRI-SENT e COP) foram considerados.

Atributos G123:

Os resultados obtidos com os atributos G123 são apresentados na Tab. 4.12.

TABELA 4.12 Resultados do Experimento 2 com atributos G123.

CF	% Acertos	Classe	Precisão	Abrangência	F-Measure
0,10	77,6%	não_correferente	**0,76**	**0,89**	**0,82**
		outra	0,80	0,62	0,70
0,35	77,1%	não_correferente	0,76	0,89	0,82
		outra	0,81	0,60	0,69

Com atributos G123, obtivemos para a classe não correferente 76% de precisão, 89% de abrangência, resultando em 82% de F-measure.

Nas duas árvores de decisão geradas, os atributos comuns são: SEM_ANTE, NUM, PRI_SENT, SUP. A segunda árvore (CF = 0,35), além desses atributos, apresenta também os atributos SA, COP, SP, TAM. O novo atributo de G123 passou a substituir vários atributos de G1 e G12 com ganho nos resultados.

Na Tab. 4.13 é apresentada a taxa de acertos e F-measure para cada combinação de atributos utilizada, destacando-se os resultados dos atributos G123.

TABELA 4.13 Percentual de acertos do Experimento 2.

Atributos	CF	% acertos	F-measure
G1	0,10	61,3%	0,64
	0,35	61,4%	0,64
G12	0,10	63,0%	0,66
	0,35	62,4%	0,65
G123	0,10	**77,6**%	**0,82**
	0,35	**77,1%**	**0,82**

Comparamos as medidas de precisão, de abrangência e de F-measure obtidas no Experimento 2 com o *baseline* (Tab. 4.14).

Com este experimento obtivemos para a classe não correferente 76% de precisão e 82% de F-measure, o que representa um ganho em relação ao *baseline* (60%, 75%). Além disso, verificaram-se ganhos significativos na taxa de acertos (de 59,8% para 77,6%).

TABELA 4.14 Comparação entre o *baseline* e Experimento 2.

Atributos	% Acertos	Classes	Precisão	Abrangência	F-measure
BASELINE	59,8%	não_correferente	**0,60**	**1,00**	**0,75**
		outra	0	0	0
G1	61,3%	não_correferente	0,71	0,58	0,64
		outra	0,53	0,66	0,59
G12	61,3%	não_correferente	0,71	0,61	0,66
		outra	0,55	0,66	0,60
G123	77,6%	não_correferente	**0,76**	**0,89**	**0,82**
		outra	0,81	0,60	0,69

Com base nos melhores resultados de classificação dos experimentos apresentados, na seção seguinte, o potencial de distinção dos atributos presentes nas árvores geradas será avaliado.

4.1.3 Potencial de Distinção das Características

De posse dos resultados da geração das árvores de decisão, os atributos presentes nas árvores com melhores resultados são analisados para avaliar o quanto estes contribuíram para a classificação das descrições definidas novas no discurso.

Desta forma, os atributos relevantes nos experimentos são analisados e expostos em forma de tabela. Cabe ressaltar que, para cada tabela a primeira coluna indica o atributo analisado; a segunda coluna apresenta o potencial de distinção do atributo isoladamente (Sozinho) e com todo o conjunto de atributos menos ele (Excluído). Já na terceira coluna está indicado o índice de precisão seguido do índice de abrangência na quarta coluna.

Experimento 1 – novas no discurso (atributos):

Para o Experimento 1 considerando a classe nova no discurso, a análise dos atributos de G1, G12 e G123 são apresentadas, respectivamente, nas tabelas: Tab. 4.15, Tab. 4.16, Tab. 4.17.

TABELA 4.15 Potencial de Distinção dos atributos G1 (CF = 0,35).

Atributos		Acertos	Precisão	Abrangência
TAM	Sozinho	**0,59**	0,67	0,33
	Excluído	0,57	0,65	0,30
SA	Sozinho	0,54	0,67	0,15
	Excluído	0,61	0,66	0,46
NP_COM	Sozinho	0,51	0,61	0,71
	Excluído	0,62	0,66	0,49
PRE_ADJ	Sozinho	0,52	0,76	0,06
	Excluído	0,62	0,65	0,53
PRE_NUM	Sozinho	0,49	0,22	0,40
	Excluído	0,63	0,66	0,54
APO_NP	Sozinho	0,50	0,58	0,38
	Excluído	0,63	0,66	0,54

TABELA 4.16 Potencial de Distinção dos atributos G12 (CF = 0,35).

Atributos		Acertos	Precisão	Abrangência
PRI_SENT	Sozinho	0,53	**1,00**	0,06
	Excluído	0,63	0,66	0,54
TAM	Sozinho	**0,59**	0,67	0,33
	Excluído	0,59	0,69	0,34
SA	Sozinho	0,54	0,67	0,15
	Excluído	0,63	0,68	0,49
PRE_ADJ	Sozinho	0,52	0,76	0,06
	Excluído	0,64	0,27	0,67
NP_COM	Sozinho	0,51	0,61	0,71
	Excluído	0,64	0,68	0,52
APO_NP	Sozinho	0,50	0,58	0,38
	Excluído	0,65	0,68	0,58

TABELA 4.17 Potencial de Distinção dos atributos G123 (CF = 0,10).

Atributos		Acertos	Precisão	Abrangência
PRI_SENT	Sozinho	0,53	**1,00**	0,06
	Excluído	0,70	0,64	0,88
SEM_ANT	Sozinho	**0,69**	0,64	0,86
	Excluído	0,54	0,92	0,89
SUP	Sozinho	0,51	0,88	0,27
	Excluído	0,70	0,64	0,87
NUM	Sozinho	0,50	0,60	0,50
	Excluído	0,70	0,65	0,88

Observamos que no Experimento 1, os maiores índices de potencial de distinção dos atributos analisados isoladamente são: TAM e SEM_ANT. Estes atributos resultaram nos menores índices da análise de todo o conjunto de atributos menos o respectivo atributo excluído, por mostrarem-se relevantes para a classificação das descrições definidas novas no

discurso. Cabe ressaltar que, o atributo PRI_SENT analisado isoladamente, apresentou o melhor índice de precisão (100%), porém possui baixa abrangência.

Com base nestes resultados, o atributo TAM destaca-se por ser um atributo original e por apresentar um resultado significativo. Analisamos a árvore de decisão utilizando os atributos G1 menos o atributo TAM para verificarmos como este atributo pode substituir outros atributos correspondentes e considerados relevantes em trabalhos relacionados. Desta forma, comparamos os resultados da classificação com todo o conjunto de atributos G1 (FIGURA 4.1) e a classificação com os atributos G1 menos o atributo TAM (G1 sem TAM) (FIGURA 4.2), ambas com CF = 0,35. Constatamos que, o atributo TAM realmente substitui os atributos SP e REL (presentes em outros trabalhos) de forma satisfatória e apresenta ganhos na taxa de acertos de 62% para 63% e de precisão de 63% para 65%. Esta comparação é mostrada na Tab 4.18.

TABELA 4.18 Comparação do atributo TAM.

Atributo	% Acertos	Classes	Precisão	Abrangência	F-measure
G1	63,0%	nova_discurso	**0,65**	0,55	0,60
		outra	0,61	0,70	0,65
G1 sem TAM	62,0%	nova_discurso	0,63	0,58	0,61
		outra	0,62	0,66	0,64

```
TAM = TRUE
|  APO_NP = TRUE: outra (3.0/1.0)
|  APO_NP = FALSE: nova_discurso (268.0/85.0)
TAM = FALSE
|  SA = TRUE: nova_discurso (94.0/32.0)
|  SA = FALSE
|  |  NP_COM = TRUE: nova_discurso (60.0/23.0)
|  |  NP_COM = FALSE
|  |  |  PRE_ADJ = TRUE: nova_discurso (24.0/6.0)
|  |  |  PRE_ADJ = FALSE
|  |  |  |  PRE_NUM = TRUE: nova_discurso (12.0/5.0)
|  |  |  |  PRE_NUM = FALSE
|  |  |  |  |  APO_NP = TRUE: nova_discurso (24.0/11.0)
|  |  |  |  |  APO_NP = FALSE: outra (620.0/229.0)
```

Figura 4.1 Árvore de Decisão com atributos G1.

```
SP = TRUE: nova_discurso (287.0/98.0)
SP = FALSE
|  SA = TRUE
|  |  PRE_ADJ = TRUE: outra (3.0/1.0)
|  |  PRE_ADJ = FALSE: nova_discurso (101.0/35.0)
|  SA = FALSE
|  |  PRE_ADJ = TRUE: nova_discurso (24.0/6.0)
|  |  PRE_ADJ = FALSE
|  |  |  NP_COM = TRUE: nova_discurso (59.0/23.0)
|  |  |  NP_COM = FALSE
|  |  |  |  REL = TRUE: nova_discurso (18.0/7.0)
|  |  |  |  REL = FALSE
|  |  |  |  |  PRE_NUM = TRUE: nova_discurso (12.0/5.0)
|  |  |  |  |  PRE_NUM = FALSE
|  |  |  |  |  |  APO_NP = TRUE: nova_discurso (25.0/11.0)
|  |  |  |  |  |  APO_NP = FALSE: outra (576.0/208.0)
```

Figura 4.2 Árvore de Decisão com atributos G1 sem TAM.

Experimento 2 – não correferentes (atributos):

Para o Experimento 2 considerando a classe não correferente, a análise dos atributos de G1, G12 e G123 são apresentados, respectivamente, nas tabelas: Tab. 4.19, Tab. 4.20, Tab. 4.21.

TABELA 4.19 Potencial de Distinção dos atributos G1 (CF = 0,35).

Atributos		Acertos	Precisão	Abrangência
TAM	Sozinho	0,58	0,58	1,00
	Excluído	0,61	0,71	0,56
NUM	Sozinho	0,58	0,58	1,00
	Excluído	0,61	0,53	0,66
APO_NP	Sozinho	0,58	0,58	1,00
	Excluído	0,60	0,70	0,55
SA	Sozinho	0,58	0,58	1,00
	Excluído	0,56	0,62	0,66
NP_COM	Sozinho	0,58	0,58	1,00
	Excluído	0,60	0,71	0,53
PRE_ADJ	Sozinho	0,58	0,58	1,00
	Excluído	0,61	0,70	0,56
SP	Sozinho	0,58	0,58	1,00
	Excluído	0,60	0,71	0,53
PRE_NUM	Sozinho	0,58	0,58	1,00
	Excluído	0,62	0,71	0,58

TABELA 4.20 Potencial de Distinção dos atributos G12 (CF = 0,35).

Atributos		Acertos	Precisão	Abrangência
PRI_SENT	Sozinho	0,58	0,58	1,00
	Excluído	0,61	0,69	0,58
TAM	Sozinho	0,58	0,58	1,00
	Excluído	0,62	0,70	0,60
APO_NP	Sozinho	0,58	0,58	1,00
	Excluído	0,61	0,71	0,58
NUM	Sozinho	0,58	0,58	1,00
	Excluído	0,62	0,71	0,61
SA	Sozinho	0,58	0,58	1,00
	Excluído	0,61	0,71	0,54
NP_COM	Sozinho	0,58	0,58	1,00
	Excluído	0,61	0,71	0,56
PRE_ADJ	Sozinho	0,58	0,58	1,00
	Excluído	0,62	0,70	0,59
SP	Sozinho	0,58	0,58	1,00
	Excluído	0,62	0,72	0,59
COP	Sozinho	0,58	0,58	1,00
	Excluído	0,63	0,72	0,60

TABELA 4.21 Potencial de Distinção dos atributos G123 (CF = 0,35).

Atributos		Acertos	Precisão	Abrangência
SEM_ANT	Sozinho	**0,76**	**0,76**	0,87
	Excluído	0,58	0,58	1,00
NUM	Sozinho	0,58	0,58	1,00
	Excluído	0,77	0,76	0,88
PRI_SENT	Sozinho	0,58	0,58	1,00
	Excluído	0,77	0,76	0,89
SUP	Sozinho	0,58	0,58	1,00
	Excluído	0,77	0,76	0,88

No Experimento 2, dos atributos analisados isoladamente, o único que apresenta o potencial de distinção é o atributo SEM_ANT. Os demais atributos isoladamente, não conseguiram distinguir os exemplos. Acreditamos que isso seja devido ao alto número de não correferentes.

4.2 Avaliação das Árvores no Ambiente ART

Nesta seção são apresentadas a implementação e a avaliação das árvores aprendidas no ambiente ART.

Para a classificação automática no ART, um conjunto de regras é aplicado aos dados de entrada (1105 exemplos do corpus 1). Cabe ressaltar que, na arquitetura da ferramenta, cada regra corresponde a um filtro (*filter*), e as informações de que filtros serão aplicados e em que seqüências, estão armazenadas em *pipes*. Dessa forma, para cada experimento foram aplicados seis *pipes* correspondentes às árvores de decisão estudadas.

Experimento 1 – novas no discurso (ART):

Os seis *pipes* do Experimento 1 e seus respectivos filtros são mostrados na Tab. 4.22.

TABELA 4.22 *Pipes* e filtros do Experimento 1.

Pipes	Filtros
pipe 1 (CF: 0,10 Atributos: G1)	TAM,SA,NP_COM,PRE_ADJ
pipe 2 (CF: 0,35 Atributos: G1)	TAM,SA,NP_COM,PRE_ADJ,PRE_NUM,APO_NP
pipe 3 (CF: 0,10 Atributos: G12)	PRI_SENT,TAM,SA,PRE_ADJ,NP_COM
pipe 4 (CF: 0,35 Atributos: G12	PRI_SENT,TAM,SA,PRE_ADJ,NP_COM,APO_NP
pipe 5 (CF: 0,10 Atributos: G123)	PRI_SENT,SEM_ANT,SUP,NUM
pipe 6 (CF 0,35 Atributos: G123)	PRI_SENT,SEM_ANT,SUP,NUM,SA,COP,SP,TAM

De posse dos resultados da aplicação de todos os *pipes* do Experimento 1, que geraram a saída no formato da ferramenta MMAX, foi realizada uma comparação entre a marcação automática e manual. Os resultados da comparação são mostrados na Tab. 4.23.

TABELA 4.23 Resultado comparativo do Experimento 1.

Pipes	Classes	Precisão	Abrangência	F-measure
pipe 1	nova_discurso	0,63	0,55	0,58
	outra	0,60	0,68	0,64
pipe 2	nova_discurso	0,63	0,58	0,60
	outra	0,61	0,66	0,64
pipe 3	nova_discurso	0,65	0,58	0,61
	outra	0,62	0,69	0,66
pipe 4	nova_discurso	**0,66**	0,60	0,63
	outra	0,64	0,69	0,66
pipe 5	nova_discurso	0,64	**0,88**	**0,74**
	outra	0,81	0,52	0,63
pipe 6	nova_discurso	0,62	0,56	0,59
	outra	0,60	0,67	0,63

Observamos que os melhores resultados para a classe nova no discurso foram: de precisão (66%) com a execução do *pipe 4;* de abrangência (88%) e de F-measure (74%) com a aplicação do *pipe 5.*

O resultado do *pipe 5* foi comparado ao resultado do *baseline*. Observamos melhorias significativas em precisão, de 50% para 64%, em F-measure de 66% para 74% e na taxa de acertos de 49,7% para 69,9% em relação ao *baseline*. Esses resultados correspondem aos resultados do aprendizado, o que valida a implementação realizada no ambiente ART.

Experimento 2 –não correferentes (ART):

Os próximos seis *pipes* do Experimento 2 e seus respectivos filtros são mostrados (Tab. 4.24) a seguir:

TABELA 4.24 *Pipes* e filtros do Experimento 2.

Pipes	Filtros
*pipe 7 (*CF: 0,10 Atributos: G1)	TAM,NUM,APO_NP,SA,NP_COM,PRE_ADJ,SP
pipe 8 (CF: 0,35 Atributos: G1)	TAM,NUM,APO_NP,SA,NP_COM,PRE_ADJ,SP, PRE_NUM
pipe 9 (CF: 0,10 Atributos: G12)	PRI_SENT,TAM,APO_NP,NUM,SA,NP_COM, PRE_ADJ,SP
pipe 10 (CF: 0,35 Atributos: G12)	PRI_SENT,TAM,APO_NP,NUM,SA,NP_COM, PRE_ADJ,SP,COP
pipe 11 (CF: 0,10 Atributos: 123)	SEM_ANT,NUM,PRI_SENT,SUP,SA,COP,SP,TAM
pipe 12 (CF: 0,35 Atributos: G123)	SEM_ANTE,NUM,PRI_SENT,SUP

Os resultados da comparação da marcação automática e manual são mostrados na Tab. 4.25.

Observamos que, com a classe não correferente os melhores resultados obtidos foram: precisão de 77% com a aplicação do *pipe 11;* 96% de abrangência e 84% de F-measure com a execução do *pipe 12.*

Um comparativo entre o resultado do *pipe 12* e do *baseline* foi realizado. Observamos que o *pipe 12* apresentou ganhos significativos em precisão (de 60% para 74%), em F-

measure (de 75% para 84%) e na taxa de acertos (de 59,8% para 78,0%) em relação ao *baseline*.

TABELA 4.25 Resultado comparativo do Experimento 2.

Pipes	Classes	Precisão	Abrangência	F-measure
pipe 7	não_correferente	0,73	0,58	0,65
	outra	0,52	0,69	0,59
pipe 8	não_correferente	0,73	0,592	0,655
	outra	0,53	0,68	0,59
pipe 9	não_correferente	0,76	0,62	0,68
	outra	0,55	0,71	0,62
pipe 10	não_correferente	0,75	0,64	0,69
	outra	0,56	0,67	0,61
pipe 11	não_correferente	**0,77**	0,83	0,80
	outra	0,72	0,62	0,67
pipe 12	não_correferente	0,74	**0,96**	**0,84**
	outra	0,89	0,51	0,65

4.2.1 Avaliação das Características em um novo corpus

Uma avaliação dessas características foi realizada em um novo corpus, o corpus Público (corpus 2). O corpus 2 difere-se do corpus 1 por estar escrito no português europeu, sendo ambos constituídos por textos jornalísticos.

Em relação aos experimentos, estes foram os mesmos aplicados ao corpus 1. O total de exemplos dessa nova base de dados (corpus 2) é apresentado na Tab. 4.26.

TABELA 4.26 Número de exemplos por classe dos Experimentos.

Experimentos	Classes adotadas	Classes	Nº Exemplos	Totais
Experimento 1	nova_discurso	nova_discurso	231	231
	outra	direta	157	252
		indireta	48	
		associativa	15	
		outra	32	
Experimento 2	não_correferente		246	246
	outra		237	237
Total de Exemplos				**483**

Conforme os experimentos apresentados na seção 4.1, serão utilizados os seguintes *baselines* para a avaliação dos resultados (Tab 4.27).

TABELA 4.27 Resultados dos *baselines* dos Experimentos.

Baselines	% Acertos	Classes	Precisão	Abrangência	F-measure
baseline1 (Experimento 1)	47,8%	nova_discurso	0,48	1,00	0,64
		outra	0	0	0
baseline2 (Experimento 2)	50,9%	não_correferente	0,51	1,00	0,67
		outra	0	0	0

Para cada experimento foram aplicados três *pipes* correspondentes aos melhores resultados da classificação automática no corpus 1 com os atributos G1, G12, G123 e correspondentes variações do fator de confiança de poda (CF). Com a execução destes seis *pipes*, foi realizada a comparação dos resultados encontrados na marcação automática e na manual.

Experimento 1 – novas no discurso Público (ART):

O resultado comparativo do Experimento 1 é mostrado na Tab. 4.28.

TABELA 4.28 Resultado comparativo do Experimento 1.

Atributos	% Acertos	Classes	Precisão	Abrangência	F-measure
baseline1	47,8%	nova_discurso	**0,48**	**1,00**	**0,64**
		outra	0	0	0
pipe 2	68,7%	nova_discurso	0,67	0,66	0,67
		outra	0,69	0,71	0,70
pipe 4	69,3%	nova_discurso	0,68	0,68	0,68
		outra	0,71	0,70	0,70
pipe 5	77,6%	nova_discurso	**0,74**	**0,81**	**0,77**
		outra	0,81	0,74	0,77

Os melhores resultados para a classe nova no discurso de precisão (74%), abrangência (81%) e F-measure (77%) foram encontrados com a aplicação do *pipe 5.* Dessa forma, para avaliar o desempenho dos atributos do *pipe 5* em relação aos demais (*pipe 2, pipe 4*), foi aplicado o teste de significância, que resultou em 99,5%.

Em todos os casos, constatamos um ganho significativo na taxa de acertos em relação ao *baseline1* com uma taxa de significância de 99,5%. Além disso, com o *pipe 5* obtivemos o melhor resultado com 74% de precisão e 77% de F-measure.

Experimento 2 – não correferentes Público (ART):

O resultado comparativo do Experimento 2 é mostrado na Tab. 4.29.

Tabela 4.29 Resultado comparativo do EXPERIMENTO 2.

Atributos	% Acertos	Classes	Precisão	Abrangência	F-measure
baseline2	50,9%	não_correferente	**0,51**	**1,00**	**0,67**
		outra	0	0	0
pipe 8	67,0%	não_correferente	0,68	0,66	0,67
		outra	0,66	0,68	0,67
pipe 10	67,5%	não_correferente	0,67	0,71	0,69
		outra	0,68	0,64	0,66
pipe 12	75,7%	não_correferente	**0,70**	**0,92**	**0,79**
		outra	0,88	0,58	0,70

Para o Experimento 2 (classe: não correferente), obtivemos precisão de 70%, abrangência de 92% e F-measure de 79% com a aplicação do *pipe 12*. O ganho do desempenho dos atributos do *pipe 12* em relação aos demais (*pipe 8* e *pipe 10*) é significante.

Nos *pipes* analisados (Tab. 4.29), observamos melhorias na taxa de acertos em relação ao *baseline2* com uma taxa de significância de 99,5%. Constatamos também que o *pipe 12* apresentou o melhor resultado em precisão (70%) e em F-measure (79%), o que representa um ganho em relação ao *baseline2*.

4.3 Considerações finais

Neste capítulo foram apresentados os experimentos realizados na ferramenta Weka e no ambiente ART.

Pode-se observar de uma forma geral que o grupo de atributos G1, que considera unicamente a estrutura do sintagma, consegue melhores resultados em relação ao *baseline* (com aumentos de precisão em torno de 15% no Experimento 1 e de 10% no Experimento 2). O grupo de atributos G12 teve uma pequena influência positiva no Experimento 2. O grupo G123 de uma forma geral supera os outros grupos, sendo que o atributo SEM_ANT substitui

os demais. O mesmo conjunto de atributos foi testado para classificações alternativas (novas no discurso e não correferente), conforme a classificação, conjuntos diferenciados de atributos foram utilizados na composição das árvores.

O resultado do aprendizado foi utilizado na implementação de árvores no ambiente ART. Com um teste realizado, nesse ambiente, com o mesmo corpus do aprendizado, obtivemos resultados correspondentes aos observados anteriormente. Isso valida a implementação proposta.

Para verificar o potencial de generalização das árvores geradas fizemos novos experimentos com um corpus novo. As árvores demonstraram um bom potencial de generalização, sendo que os resultados apresentaram taxas de acerto em torno de 75% comparáveis a taxa de acertos encontrados para o corpus 1.

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou um estudo de características das descrições definidas que pode auxiliar na identificação das expressões novas no discurso e das expressões não correferentes. Este estudo é importante, pois o número de expressões sem antecedentes textuais no discurso é expressivo. Por exemplo, no corpus da Folha de São Paulo (seção 3.1), das 1105 descrições definidas analisadas, 550 são novas no discurso e 661 são não correferentes de acordo com a marcação manual (seção 3.3).

Desta forma, um estudo de corpus foi realizado para investigar as características significativas das descrições definidas da Língua Portuguesa com base na estrutura do sintagma, análise da sentença e do texto. Partimos de estudos realizados em outras línguas apresentados na seção 2.5, onde foram analisadas quais características da Língua Inglesa são válidas para a Língua Portuguesa. Além dessas, outras características foram adicionadas e organizadas em grupos na seção 3.7.

De posse dos grupos de características, aplicou-se o algoritmo de árvores de decisão *j48* (seção 3.5) para a investigação da relevância dessas características para a classificação das expressões novas no discurso e não correferentes.

Com a execução dos experimentos na ferramenta Weka (seção 4.1), foram geradas árvores de decisão e os atributos que apresentaram os melhores resultados foram analisados.

Como resultados, a combinação de atributos G123 apresentou os melhores índices de desempenho no Experimento 1 (precisão de 65%, abrangência de 88%, F-measure de 75%) e no Experimento 2 (precisão de 76%, abrangência de 89%, F-measure de 82%). Nas duas árvores de decisão correspondentes, são comuns os atributos: PRI_SENT, SEM_ANT, SUP, NUM. No Experimento 2 (classe: não correferente), além desses atributos, também

apresentam os atributos: SA, COP, SP, TAM. A taxa de acertos aumentou de 18% a 20% nesses experimentos.

As árvores de decisão geradas com a ferramenta Weka foram implementadas e avaliadas no ambiente ART em um novo corpus, com textos do jornal português Público. Como resultados, obtivemos para o *pipe 5* (atributos: G123, classe: nova no discurso) um índice de 74% de precisão, 81% de abrangência e 77% de F-measure. Na comparação com o *baseline1* foi constatado um aumento de 29,8% na taxa de acerto (de 47,8% para 77,6%) com grau de significância de 99,5%. É interessante observar que o *pipe 2*, com os atributos G1 baseados apenas na estrutura do sintagma, já apresenta melhoras em relação ao *baseline1* (de 20,9% na taxa de acertos).

Com o *pipe 12* (atributos: G123, classe: não correferente) obtivemos 70% de precisão, 92% de abrangência e 79% de F-measure. Da comparação com seu *baseline2* foi constatado um aumento de 24,8% na taxa de acerto (de 50,9% para 75,7%) com grau de significância de 99,5%. Cabe ressaltar que, o *pipe 8* com os atributos G1 (baseados apenas na estrutura do sintagma) já apresenta ganho em relação ao *baseline2* (de 16,1% na taxa de acertos).

Os trabalhos relacionados (seção 2.6) reportam resultados da classificação de expressões referenciais com base em outros corpora e línguas, o que torna a comparação difícil. No entanto, apresentamos aqui alguns desses resultados para dar uma idéia dos níveis alcançados para essa tarefa em outros estudos (Tab 5.1). Cabe ressaltar que, neste trabalho diferentemente dos demais, foi realizada uma análise detalhada dos atributos utilizados na classificação, contribuindo para o entendimento do problema que está sendo analisado e que a distinção entre novas no discurso e não correferentes foi abordada.

Como trabalhos futuros, pretendemos realizar uma análise detalhada dos erros ocorridos na classificação das descrições definidas e dar continuidade a análise de atributos relevantes para a classificação em questão. Outro objetivo é realizar uma investigação das

demais classificações das descrições definidas (anafórica direta, indireta, associativa). Na análise das descrições definidas anafóricas associativas analisaremos as marcações semânticas fornecidas pelo analisador sintático PALAVRAS. Pretende-se, também associar este trabalho a um estudo de correferência pronominal, o qual já foi iniciado com primeiros resultados em AIRES et al. (2004).

Além disso, pretendemos aplicar uma extensão deste estudo das descrições definidas, no sistema de perguntas e respostas sobre textos jurídicos da Procuradoria Geral da República de Portugal em desenvolvimento junto ao Departamento de Informática da Universidade de Évora.

TABELA 5.1 Resultados dos trabalhos relacionados.

Trabalhos Relacionados	F*measure*	Abordagem
AONE, BENNET (1995)	0,77	resolução de anáforas definidas e pronominais
MCCARTHY, LEHNERT (1995)	0,86	resolução de correferência de sintagmas nominais
BEAN, RILOFF (1999)	0,82	classificação de descrições definidas não anafóricos
CARDIE, WAGSTAFF (1999)	0,53	resolução de correferência de sintagmas nominais
VIEIRA, POESIO (2000)	0,70	classificação de descrições definidas novas no discurso
SOON; NG; LIM (2001)	0,62	resolução de correferência de sintagmas nominais
NG, CARDIE (2002a)	0,66	classificação de sintagmas nominais anafóricos e não anafóricos
MULLER; RAPP; STRUBE (2002)	0,80	resolução de pronomes e descrições definidas
STRUBE; RAPP; MULLER (2003)	0,68	resolução de correferência de pronomes e descrições definidas
URYUPINA (2003)	0,83	classificação de entidades novas no discurso e únicas
POESIO et al (2005)	0,82	classificação de descrições definidas novas no discurso
COLLOVINI (2005) no Weka com o corpus 1	0,75	classificação de descrições definidas novas no discurso
COLLOVINI (2005) no weka com o corpus 1	0,82	classificação de descrições definidas não correferentes
COLLOVINI (2005) no ART com o corpus 1	0,74	classificação de descrições definidas novas no discurso
COLLOVINI (2005) no ART com o corpus 1	0,84	classificação de descrições definidas não correferentes
COLLOVINI (2005) no ART com o corpus 2	0,77	classificação de descrições definidas novas no discurso
COLLOVINI (2005) no ART com o corpus 2	0,79	classificação de descrições definidas não correferentes

Referências

AIRES, Ana Margarida et al. **Avaliação de Centering em Resolução Pronominal da Língua Portuguesa.** In: Workshoop Herramientas y Recursos Linguísticos para el Español y el Portugués (IBERAMIA). Tonantzintla, México, 22 - 26 de novembro de 2004.

AMADO, Nuno Manuel Reis. **Algoritmos Paralelos de Indução de Árvores de Decisão**. Dissertação (Mestrado em Engenharia) - Faculdade de Engenharia, Universidade de Porto, Porto, 2001.

AONE, Chinatsu; BENNETT, Scott W**. Evaluating automated and manual acquisition of anaphora resolution strategies.** In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, Massachusetts, 26–30 June 1995, p. 122– 129.

BEAN, David L.; RILOFF, Ellen. **Corpus-based Identification of Non-Anaphoric Noun Phrases**. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999). University of Maryland, College Park, Maryland, USA, 20-26 June 1999, p. 373–380.

BICK, Eckhard. **The Parsing System PALAVRAS:** Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Tese (Doutorado) - Arhus University, Arhus, 2000.

BONINI, Adair. **Gêneros textuais e cognição:** um estudo sobre a organização cognitiva da identidade dos textos. Florianópolis: Insular, 2002, 239 p.

BREIMAN, Leo et al. **Classification and Regression Trees**. Wadsworth and Brooks, Monterey, Ca, 1984.

CARDIE, Claire; WAGSTAFF, Kiri. **Noun phrase coreference as clustering**. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (SIGDAT 1999). College Park, Md., 21–22 June 1999, p. 82–89.

COHEN, William W. **Fast Effective Rule Induction**. In: Proceedings of the 12^{TH} International Conference on Machine Learning. Lake Tahoe, California, 1995, p. 115-123.

COLLOVINI, Sandra; GOULART, Rodrigo; VIEIRA, Renata. **Identificação de Expressões Anafóricas e Não Anafóricas com Base na Estrutura do Sintagma**. In: 2.º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2004). Salvador, Bahia, 05 e 06 de agosto de 2004.

FÁVERO, Leonor Lopes. **Coesão e Coerência Textuais**. 4. ed. São Paulo: Ática, 1997. 104 p.

FÁVERO, Leonor Lopes; KOCH, Ingedore Grunfeld Villaça. **Lingüística** textual**:** introdução. 5. ed. São Paulo: Cortez, 1994, 105 p.

GAMMA, Erich. **Design Patterns:** Elements of Reusable Object-Oriented Software. Reading: Addison-Wesley, 1995, 395 p.

GASPERIN, Caroline et al. *Extrating XML* **Syntactic Chunks from Portuguese Corpora**. In: Traitement Automatique Dês Langues Minoritaires(TALN 2003). Btaz-sur-mer, France, 2003.

GASPERIN, Caroline; GOULART, Rodrigo; VIEIRA, Renata. **Uma Ferramenta para Resolução Automática de Correferência**. In: Anais do Encontro Nacional de Inteligência Artificial (ENIA 2003). Campinas, SP, 2003.

GOULART, Rodrigo; GASPERIN, Carolina; VIEIRA, Renata. **Uma Ferramenta para Resolução Automática de Correferência**. Scientia, v. 14, n. 2, São Leopoldo, 2004, p. 238-256.

HAUSSER, Roland. **Foundations of computational linguistics***:* man-machine communication in natural language. Erlangen: Springer, 1999, 534 p.

JURAFSKY, Daniel; MARTIN, James H. **Speech and language processing***:* an introduction to natural language processing, computational linguistics and speech recognition. Upper Saddle River: Prentice-Hall, 2000, 934 p.

KOCH, Ingedore Grunfeld Villaça. **Desvendando os segredos do texto**. 2. ed. São Paulo: Cortez, 2003. 168 p.

KOCH, Ingedore Grunfeld Villaça. **Argumentação e linguagem**. 6. ed. São Paulo: Cortez, 2000, 240 p.

KOCH, Ingedore Grunfeld Villaça; TRAVAGLIA, Luiz Carlos. **A coerência textual**. São Paulo: Contexto, 2002, 94 p.

KOCH, Ingedore Grunfeld Villaça; TRAVAGLIA, Luiz Carlos. **A coerência textual**. 7. ed. São Paulo: Contexto, 1996, 94 p.

KORFHAGE, Robert R. **Information Retrieval and Storage**. New York: John Wiley & Sons, 1997, 349 p.

KOWALSKI, Gerald. **Information Retrieval Systems: Theory and Implementation**. Boston: Kluwer Academic Publishers, 1997, 282 p.

MACAMBIRA, José Rebouças. **A estrutura morfo-sintática do português***:* aplicação do estruturalismo lingüístico. 6. ed. São Paulo: Pioneira, 1990, 363 p.

MCCARTHY, Joseph F.; LEHNERT, Wendy G. **Using decision trees for coreference resolution.** In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, Canada, 1995, p. 1050–1055.

MITCHELL, Tom Michael. **Machine learning**. Boston: McGraw-Hill, 1997, 414 p.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. **Indução de Árvores de Decisão**, Capítulo 5, Solange Oliveira Rezende. In: **Sistemas Inteligentes**: Fundamentos e Aplicações. Baueri, SP, Manole, 2003, p. 115-139.

MUC-6. **Coreference task definition**. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), v. 2.3. San Francisco, CA, 8 Sep 1995, p. 335-344.

MUC-7**. Coreference task definition**. In: Proceedings of the Seventh Message Understanding Conference (MUC-7), v. 3.0. San Francisco, CA, 13 Jul 1997.

MULLER, Christoph; STRUBE ,Michael. **MMAX:** A tool for the annotation of multi-modal corpora. In: Proceedings of the IJCAI 2001. Seattle, 2000, p. 45–50.

MULLER, Christoph; RAPP, Stefan; STRUBE, Michael. **Applying Co-training to reference resolution**. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). Philadelphia, Penn., 7–12 July 2002, p. 352-359.

NG, Vincent; CARDIE, Claire. **Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution**. In: Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan, 2002, p. 730–736.

NG, Vincent; CARDIE, Claire. **Improving machine learning approaches to coreference resolution**. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, 2002.

POESIO, Máximo et al. **Does Discourse-new Detection Help Definite Description Resolution?** In: Sixth International Workshop on Computational Semantics (IWCS 6). Tiburg, 2005.

PRINCE, Ellen F. **Toward Taxonomy of given-new information**. In: P. Cole, editor, Radical Gramatics. Academic Press, New York, 1981, p. 223-256.

QUINLAN, John Ross. **Induction of Decision Trees**. Machine Learning, v. 1, n. 1, 1986, p. 81 – 106.

QUINLAN, John Ross. **C4.5:** programs for machine learning. San Mateo: Morgan Kaufmann, 1993. 302 p.

SANTOS, Diana. **O projecto Processamento Computacional do Português: Balanço e Perspectivas**. In: V Encontro para o Processamento da Língua Portuguesa Escrita e Falada (PROPOR). Atibaia, São Paulo, Brasil, 2000, p. 105-113.

SANTOS, Diana. **Introdução ao Processamento da Linguagem Natural através das Aplicações.** In: Ranchhod. Tratamento das Línguas pro Computador – Uma introdução Lingüística Computacional e suas Aplicações. Lisboa, Caminho, 2001, p. 229-259.

SOON, Wee Meng; NG, Hwee Tou; LIM, Daniel Chung Yong. **A machine learning approach to coreference resolution of noun phrases**. In: Computational Linguistics, v. 27, n. 4, 2001, p. 521–544.

STRUBE, Michael; RAPP, Stefan; MULLER, Christoph. **The influence of minimum edit distance on reference resolution**. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. Philadelphia, Pa., 6–7 July 2002, p. 312-319.

SALMON-ALT, Susanne; VIEIRA, Renata. **Nominal Expressions in Multilingual Corpora: Definites and Demonstratives**. In: Preceedings of teh LREC. Lãs Palmas de Gran Canária, 2002.

URYUPINA, Olga. **High-precision Identification of Discourse New and Unique Noun Phrases.** In: Proceedings of the ACL Student Workshop. Sapporo, 2003.

VIEIRA, Renata et al. **From concrete to virtual annotation mark-up language: the case of COMMOn-REFs.** In: Proceedings of the ACL 2003 - Workshop on Linguistic Annotation: Getting the Model Right. Sapporo, 2003.

VIEIRA, Renata; GASPERIN, Caroline; GOULART, Rodrigo. **From manual to automatic annotation of coreference**. In: Proceedings of the International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization. Venice, 2003.

VIEIRA, Renata; SALMON-ALT, Susanne; SCHANG, Emmanuel. **Multilingual corpora annotation for processing definite descriptions**. In: Proceedings of the PorTAL. Faro, Portugal, 2002, p. 249-258.

VIEIRA, Renata et al. **Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus**. In: Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium. Lisboa, 2002. Edições Colibri, v.1, p. 233 – 238.

VIEIRA, Renata; POESIO, Massimo. **An empirically-based system for processing definite descriptions**. In: Computational Linguistics, v. 26, n. 4, 2000, p. 539–594.

VIEIRA, Renata. **Definite description processing in unrestricted text**. Tese (Doutorado) - University of Edinburgh, Edinburgh, 1998.

VILELA, Mário Augusto do Quinteiro; KOCH, Ingedore Grunfeld Villaça. **Gramática da Língua Portuguesa:** gramática da palavra, gramática da frase, gramática do texto/discurso. Coimbra: Almedina, 2001, 565 p.

WITTEN, Ian H.; FRANK, Eibe. **Data mining:** practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann, 2000, 369 p.

ZUMTHOR, Paul. **Performance, recepção, leitura**. São Paulo: Educ, 2000, 137 p.

APÊNDICE A – Folhas de Estilo XSL

A seguir é apresentada a folha de estilos em XSL para a extração das descrições definidas de cada texto do corpus.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!--DESCRICAO: Extração das Descrições definidas-->

<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="xml" encoding="ISO-8859-1" doctype-
system="anaphor.dtd"/>
    <xsl:param name="fwords" select="../data/default_words.xml"/>
    <xsl:param name="fpos" select="../data/default_pos.xml"/>
    <xsl:param name="fchunks" select="../data/default_chunks.xml" />
    <xsl:param name='debug' select="true" />
    <xsl:variable name='words' select="document($fwords)"/>
    <xsl:variable name='pos' select="document($fpos)"/>
    <xsl:variable name='chunks' select="document($fchunks)"/>
      <xsl:include
href="http://www.inf.unisinos.br/~renata/tools/write_words.xsl"/>

      <xsl:template match="/text">
            <xsl:element name="anaphor-set">
                  <xsl:apply-templates select="paragraph"/>
            </xsl:element>
      </xsl:template>
      <xsl:template match="paragraph">
            <xsl:apply-templates select="sentence"/>
      </xsl:template>

      <xsl:template match="sentence">
            <xsl:apply-templates select="chunk"/>
      </xsl:template>

      <xsl:template match="chunk">
            <xsl:if test="@form='np'">
                  <xsl:variable name="first_word" select="substring-
before(@span,'..')"/>
                        <xsl:if
test="$pos//word[@id=$first_word]/art/secondary_art/@tag = 'artd'">
                        <xsl:element name="anaphor">
                              <xsl:attribute name="span">
                                    <xsl:value-of select="@span"/>
                              </xsl:attribute>
                              <xsl:attribute name="pointer">
                                    <xsl:value-of select="''"/>
                              </xsl:attribute>

                              <xsl:if test="$debug='true'">
                                    <xsl:call-template name="write_words">
                                          <xsl:with-param name="span">
```

```
                                        <xsl:value-of select="@span"/>
                                      </xsl:with-param>
                                    </xsl:call-template>
                                </xsl:if>

                            </xsl:element>
                            <xsl:value-of select="'&#010;'"/>
                        </xsl:if>

                    </xsl:if>
                    <xsl:apply-templates select="chunk"/>
                </xsl:template>

</xsl:stylesheet>
```

A seguir é apresentada a folha de estilos em XSL para a identificação automática das características 1,5,6 nas descrições definidas de cada texto do corpus.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- DESCRICAO: Identificação automática das características 1,4,6 nas
descrições definidas -->

<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="xml" encoding="ISO-8859-1"/>
<xsl:include href="write_words.xsl"/>
      <xsl:param name="fwords" select="../data/default_words.xml"/>
      <xsl:param name="fpos" select="../data/default_pos.xml"/>
      <xsl:param name="fchunks" select="../data/default_chunks.xml"/>
      <xsl:param name="fcandidates" select="../data/candidates.xml"/>
      <xsl:param name="debug" select="'true'"/>
      <xsl:param name="ap" select="'true'"/>
      <xsl:param name="pp" select="'true'"/>
      <xsl:param name="prop_comp" select="'true'"/>
      <xsl:variable name="words" select="document($fwords)"/>
      <xsl:variable name="pos" select="document($fpos)"/>
      <xsl:variable name="chunks" select="document($fchunks)"/>
      <xsl:variable name="candidates" select="document($fcandidates)"/>

      <xsl:template match="/anaphor-set">
            <xsl:element name="anaphor-set"><xsl:value-of
select="'&#010;'"/>
                  <xsl:apply-templates select="anaphor"/>
            </xsl:element>
      </xsl:template>

      <xsl:template match="anaphor">
            <xsl:element name="anaphor">
                  <xsl:variable name="v_anaphor_span" select="@span"/>
                  <xsl:variable name="v_pp"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@form='pp']/@span"/>
                  <xsl:variable name="v_ap"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@form='ap']/@span"/>
                  <xsl:variable name="v_prop_comp"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h' and
@form='prop']/@span"/>
                  <xsl:attribute name="span">
                        <xsl:value-of select="@span"/>
                  </xsl:attribute>
                  <xsl:attribute name="pointer">
                        <xsl:value-of select="@pointer"/>
                  </xsl:attribute>
                  <xsl:attribute name="classification">

                  <!-- DESCRICAO: identificação da característica 1 -->

                        <xsl:if test="$v_pp != ''">
                              <xsl:value-of select="'nova_discurso'"/>
                        </xsl:if>

                  <!-- DESCRICAO: identificação da característica 4 -->

                        <xsl:if
test="contains($words//word[@id=$v_prop_comp],'_')">
                              <xsl:value-of select="'nova_discurso'"/>
```

```
                        </xsl:if>

                    <!-- DESCRICAO: identificação da característica 6 -->

                        <xsl:if test="$v_ap != ''">
                              <xsl:value-of select="'nova_discurso'"/>
                        </xsl:if>

                    </xsl:attribute>
                    <xsl:if test="$debug='true'">
                        <xsl:call-template name="write_words">
                              <xsl:with-param name="span">
                                    <xsl:value-of select="@span"/>
                              </xsl:with-param>
                        </xsl:call-template>
                    </xsl:if>
                </xsl:element><xsl:text>
            </xsl:text>
            </xsl:template>

</xsl:stylesheet>
```

A seguir é apresentada a folha de estilos em XSL para gerar o script de entrada para a Ferramenta Weka, considerando as classes: nova no discurso e outra.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- DESCRICAO: saida com os vetores-exemplo das DDs com presença/ausência
de suas caracteristicas e a respectiva classe associada-->
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="text" encoding="ISO-8859-1" doctype-
system="anaphor.dtd"/>
<xsl:include href="write_words.xsl"/>
 <xsl:param name="fwords" select="../data/default_words.xml"/>
 <xsl:param name="fpos" select="../data/default_pos.xml"/>
 <xsl:param name="fchunks" select="../data/default_chunks.xml"/>
 <xsl:param name="fcandidates" select="../data/candidates.xml"/>
 <xsl:param name="fspecial_words" select="special_words.xml"/>
 <xsl:param name="debug" select="'true'"/>
 <xsl:variable name="v-found" select="false"/>
 <xsl:variable name="words" select="document($fwords)"/>
 <xsl:variable name="pos" select="document($fpos)"/>
 <xsl:variable name="chunks" select="document($fchunks)"/>
 <xsl:variable name="candidates" select="document($fcandidates)"/>
 <xsl:variable name="special_words" select="document($fspecial_words)"/>
<xsl:template match="/markables">&#64;relation script
<!-- ATRIBUTOS-->
&#64;attribute S_ANT {TRUE,FALSE}
&#64;attribute PRE_ADJ {TRUE,FALSE}
&#64;attribute PRE_NUM {TRUE,FALSE}
&#64;attribute NUM {TRUE,FALSE}
&#64;attribute REL {TRUE,FALSE}
&#64;attribute NP_COM {TRUE,FALSE}
&#64;attribute COP {TRUE,FALSE}
&#64;attribute SP {TRUE,FALSE}
&#64;attribute SA {TRUE,FALSE}
&#64;attribute APO {TRUE,FALSE}
&#64;attribute APO_NP {TRUE,FALSE}
&#64;attribute SUP {TRUE,FALSE}
&#64;attribute SUP_A {TRUE,FALSE}
&#64;attribute PRI_SENT {TRUE,FALSE}
&#64;attribute TAM {TRUE,FALSE}
&#64;attribute DET {TRUE,FALSE}
&#64;attribute category {nova_discurso,outra}
&#64;data
<xsl:apply-templates select="markable[@form='defnp' and @span =
$chunks//chunk/@span]"/>
</xsl:template>
 <xsl:template match="markable">
  <xsl:variable name="v_anaphor_span" select="@span"/>

            <!-- BUSCA ANTECEDENTE -->

            <!-- Head Chunk span which has a father with the same anaphor's
span  -->
            <xsl:variable name="v-chunk-span"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/@span"/>


            <!-- Anaphor's head word -->
            <xsl:variable name="v-anaphor-word-pos" select="substring-
after($v-chunk-span,'_')"/>
```

```
            <xsl:variable name="v-anaphor-word"
select="$words//word[@id=$v-chunk-span]"/>


            <!-- Test each word-->
            <xsl:variable name="v-anaphor-antecedent">
                  <xsl:for-each select="$words//word[substring-
after(@id,'_') &lt; substring-after($v-chunk-span,'_')]">
                        <xsl:if test=". = $v-anaphor-word">
                              <xsl:value-of select="$words//word[@id =
$v-chunk-span]"/> ,
                        </xsl:if>
                  </xsl:for-each>
            </xsl:variable>

            <!-- CRIACAO VETOR-->
            <xsl:choose>
                  <xsl:when test="$v-anaphor-antecedent =
''">TRUE,</xsl:when>
                  <xsl:otherwise>FALSE,</xsl:otherwise>
            </xsl:choose>

  <!-- PRE-MODIFICADOR ADJETIVO-->
  <xsl:variable name="v_pre-adj1" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/preceding-
sibling::chunk/@span]/adj/@canon"/>
  <!-- CASOS EM QUE O PRE-MODIFICADOR ESTA COMO VERBO NO PARTICIPIO E É
ADJETIVO-->
  <xsl:variable name="v_pre-adj2" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/preceding-
sibling::chunk/@span]/v/pcp/@number"/>

  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_pre-adj1!='' or $v_pre-adj2!=''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>

  <!-- PRE-MODIFICADOR NUMERAL-->
               <xsl:variable name="v_pre-num" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/preceding-
sibling::chunk/@span]/num/@canon"/>

  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_pre-num != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>

  <!-- POS-MODIFICADOR NUMERAL-->
               <xsl:variable name="v_pos-num" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/following-
sibling::chunk/@span]/num/@canon"/>

  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_pos-num != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>

  <!-- CLAUSULA RELATIVA-->
```

```
  <xsl:variable name="v_np"
select="$chunks//chunk[@span=$v_anaphor_span]"/>
  <!-- Verificar se possui um antecedente com a forma de pronome relativo -
->
  <xsl:variable name="v_relativa"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk/chunk[@form='pron_indp'
or @form='adv' or @form='pron_det']/@span"/>

  <xsl:variable name="v_relativa_teste" select="$special_words//group[word
= $words//word[@id = $v_relativa]]/@name"/>

  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_relativa_teste = 'Relative
pronouns'">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>


  <!-- NUCLEO NOME PROPRIO COMPOSTO SEM "_" -->
               <xsl:variable name="v_prop_comp1" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h' and
@form='prop']/following-sibling::chunk/@span]/prop/@canon"/>
  <!-- NUCLEO NOME PROPRIO COMPOSTO COM "_" -->
  <xsl:variable name="v_prop_comp"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h' and
@form='prop']/@span"/>

  <!-- CRIACAO VETOR-->
        <xsl:choose>
   <xsl:when test="contains($words//word[@id=$v_prop_comp],'_') or
$v_prop_comp1 != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>

  <!-- COPULAR-->
  <!--verificar o primeiro argumento da construcao copular-->

  <xsl:variable name="v_cop" select="$chunks//chunk[@span=$v_anaphor_span
and @ext='subj']/@span"/>
   <xsl:variable name="v_copula" select="substring-after(substring-
after($v_cop,'..'),'_')"/>
   <xsl:variable name="v_t" select="concat('word','_',$v_copula+1)"/>
   <xsl:variable name="v_teste" select="$pos//word[@id=$v_t]/v/@canon"/>

  <!--verificar o segundo argumento da construcao copular-->

  <xsl:variable name="v_copular"
select="$chunks//chunk[@span=$v_anaphor_span and @ext = 'sc']/@span"/>

      <!-- CRIACAO VETOR-->
      <xsl:choose>
   <xsl:when test="$v_teste='ser' or $v_teste='estar' or $v_copular!=
''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
     </xsl:choose>

  <!-- POS MODIFICADOR SINTAGMA PREPOSICIONAL -->
  <xsl:variable name="v_pp"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@form='pp']/@span"/>
```

```
  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_pp != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>

  <!-- POS MODIFICADOR SINTAGMA ADJETIVAL -->
            <xsl:variable name="v_Sadj"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@form='ap']/@span"/>
  <!-- POS-MODIFICADOR ADJETIVO-->
            <xsl:variable name="v_pos-adj_1" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/following-
sibling::chunk/@span]/adj/@canon"/>
  <!-- POS-MODIFICADOR SENDO VERBO NO PARTICIPIO E É ADJETIVO-->
            <xsl:variable name="v_pos-adj_2" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/following-
sibling::chunk/@span]/v/pcp/@number"/>

     <!-- CRIACAO VETOR-->
     <xsl:choose>
   <xsl:when test="$v_Sadj != '' or $v_pos-adj_1 != '' or $v_pos-adj_2 !=
''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
     </xsl:choose>

  <!-- APOSTO DE FORMA EXPLICITA-->
  <xsl:variable name="v_app"
select="$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='app']/@span"/>

  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_app != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>

  <!-- APOSTO NOME PROPRIO-->
               <xsl:variable name="v_app-prop" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/following-
sibling::chunk/@span]/prop/@canon"/>

  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_app-prop != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
  </xsl:choose>

  <!-- SUPERLATIVO-->
            <xsl:variable name="v_super" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@ext='h']/preceding-
sibling::chunk/@span]/adj/secondary_adj/@tag"/>

     <!-- CRIACAO VETOR-->
     <xsl:choose>
   <xsl:when test="$v_super != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
     </xsl:choose>

  <!-- SUPERLATIVO ABSOLUTO -->
            <xsl:variable name="v_adj" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span and @ext='sc']/chunk[@ext='h' and
@form='adj']/@span]/adj/secondary_adj/@tag"/>
```

```
     <xsl:variable name="v_tamanho_abs" select="substring-after(substring-
after($v_anaphor_span,'..'),'_') - substring-after(substring-
before($v_anaphor_span,'..'),'_') "/>

  <!-- CRIACAO VETOR-->
            <xsl:choose>
   <xsl:when test="$v_adj != '' and $v_tamanho_abs = 1">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
     </xsl:choose>

     <!-- PRIMEIRA SENTENÇA-->
            <xsl:variable name="v_pri-sent"
select="$chunks/text[1]/paragraph[1]/sentence[1]/descendant::chunk[@span =
$v_anaphor_span]/@span"/>

  <!-- CRIACAO VETOR-->
  <xsl:choose>
   <xsl:when test="$v_pri-sent != ''">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
     </xsl:choose>

     <!-- TAMANHO-->
            <xsl:variable name="v_tamanho" select="substring-
after(substring-after($v_anaphor_span,'..'),'_') - substring-
after(substring-before($v_anaphor_span,'..'),'_') "/>

            <!-- CRIACAO VETOR-->
     <xsl:choose>
   <xsl:when test="$v_tamanho >= 4">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
     </xsl:choose>

     <!-- DDEF COMPLEXA-->
            <xsl:variable name="v_pron" select="$pos//word[@id =
$chunks//chunk[@span=$v_anaphor_span]/chunk[@form='pron_det']/@span]/pron/s
econdary_pron/@tag"/>

     <!-- CRIACAO VETOR-->
     <xsl:choose>
   <xsl:when test="$v_pron = 'poss' or $v_pron = 'dem' or $v_pron =
'quant'">TRUE,</xsl:when>
   <xsl:otherwise>FALSE,</xsl:otherwise>
     </xsl:choose>

     <!-- CLASSIFICACAO-->
     <xsl:choose>
   <xsl:when
test="@classification='discourse_new'">nova_discurso</xsl:when>
   <xsl:otherwise>outra</xsl:otherwise>
     </xsl:choose>


 <xsl:text>
  </xsl:text>
  </xsl:template>

</xsl:stylesheet>
```

APÊNDICE B – Árvores de Decisão

EXPERIMENTO 1

Árvore de Decisão do Experimento 1 com atributos G1 e CF 0,10

```
TAM = TRUE: nova_discurso (271.0/87.0)
TAM = FALSE
|  SA = TRUE: nova_discurso (94.0/32.0)
|  SA = FALSE
|  |  NP_COM = TRUE: nova_discurso (60.0/23.0)
|  |  NP_COM = FALSE
|  |  |  PRE_ADJ = TRUE: nova_discurso (24.0/6.0)
|  |  |  PRE_ADJ = FALSE: outra (656.0/249.0)
```

Árvore de Decisão do Experimento 1 atributos G1 e CF 0,35

```
TAM = TRUE
|  APO_NP = TRUE: outra (3.0/1.0)
|  APO_NP = FALSE: nova_discurso (268.0/85.0)
TAM = FALSE
|  SA = TRUE: nova_discurso (94.0/32.0)
|  SA = FALSE
|  |  NP_COM = TRUE: nova_discurso (60.0/23.0)
|  |  NP_COM = FALSE
|  |  |  PRE_ADJ = TRUE: nova_discurso (24.0/6.0)
|  |  |  PRE_ADJ = FALSE
|  |  |  |  PRE_NUM = TRUE: nova_discurso (12.0/5.0)
|  |  |  |  PRE_NUM = FALSE
|  |  |  |  |  APO_NP = TRUE: nova_discurso (24.0/11.0)
|  |  |  |  |  APO_NP = FALSE: outra (620.0/229.0)
```

Árvore de Decisão do EXPERIMENTO 1 atributos G12 e CF 0,10

```
PRI_SENT = TRUE: nova_discurso (35.0)
PRI_SENT = FALSE
|  TAM = TRUE: nova_discurso (261.0/87.0)
|  TAM = FALSE
|  |  SA = TRUE: nova_discurso (91.0/32.0)
|  |  SA = FALSE
|  |  |  PRE_ADJ = TRUE: nova_discurso (24.0/6.0)
|  |  |  PRE_ADJ = FALSE
|  |  |  |  NP_COM = TRUE: nova_discurso (56.0/23.0)
|  |  |  |  NP_COM = FALSE: outra (638.0/231.0)
```

Árvore de Decisão do EXPERIMENTO 1 com atributos G12 e CF 0,35

```
PRI_SENT = TRUE: nova_discurso (35.0)
PRI_SENT = FALSE
|  TAM = TRUE
|  |  APO_NP = TRUE: outra (3.0/1.0)
|  |  APO_NP = FALSE: nova_discurso (258.0/85.0)
|  TAM = FALSE
|  |  SA = TRUE: nova_discurso (91.0/32.0)
|  |  SA = FALSE
|  |  |  PRE_ADJ = TRUE: nova_discurso (24.0/6.0)
|  |  |  PRE_ADJ = FALSE
|  |  |  |  NP_COM = TRUE: nova_discurso (56.0/23.0)
|  |  |  |  NP_COM = FALSE
|  |  |  |  |  APO_NP = TRUE: nova_discurso (24.0/11.0)
|  |  |  |  |  APO_NP = FALSE: outra (614.0/218.0)
```

Árvore de Decisão do EXPERIMENTO 1 com atributos G123 e CF 0,10

```
PRI_SENT = TRUE: nova_discurso (35.0)
PRI_SENT = FALSE
|  SEM_ANT = TRUE: nova_discurso (705.0/260.0)
|  SEM_ANT = FALSE
|  |  SUP = TRUE: nova_discurso (6.0/1.0)
|  |  SUP = FALSE
|  |  |  NUM = TRUE: nova_discurso (4.0/1.0)
|  |  |  NUM = FALSE: outra (355.0/62.0)
```

Árvore de Decisão do EXPERIMENTO 1 com atributos G123 e CF 0,35

```
PRI_SENT = TRUE: nova_discurso (35.0)
PRI_SENT = FALSE
|  SEM_ANT = TRUE: nova_discurso (705.0/260.0)
|  SEM_ANT = FALSE
|  |  SUP = TRUE: nova_discurso (6.0/1.0)
|  |  SUP = FALSE
|  |  |  NUM = TRUE: nova_discurso (4.0/1.0)
|  |  |  NUM = FALSE
|  |  |  |  SA = TRUE
|  |  |  |  |  COP = TRUE
|  |  |  |  |  |  SP = TRUE: outra (3.0/1.0)
|  |  |  |  |  |  SP = FALSE: nova_discurso (4.0/1.0)
|  |  |  |  |  COP = FALSE
|  |  |  |  |  |  TAM = TRUE: nova_discurso (4.0/1.0)
|  |  |  |  |  |  TAM = FALSE: outra (24.0/9.0)
|  |  |  |  SA = FALSE: outra (320.0/46.0)
```

EXPERIMENTO 2

Árvore de Decisão do EXPERIMENTO 2 com atributos G1 e CF 0,10

```
TAM = TRUE: nao_correferente (271.0/70.0)
TAM = FALSE
|  NUM = TRUE: nao_correferente (6.0/1.0)
|  NUM = FALSE
|  |  APO_NP = TRUE: nao_correferente (33.0/8.0)
|  |  APO_NP = FALSE
|  |  |  SA = TRUE: nao_correferente (92.0/29.0)
|  |  |  SA = FALSE
|  |  |  |  NP_COM = TRUE: nao_correferente (55.0/19.0)
|  |  |  |  NP_COM = FALSE
|  |  |  |  |  PRE_ADJ = TRUE: nao_correferente (21.0/4.0)
|  |  |  |  |  PRE_ADJ = FALSE
|  |  |  |  |  |  SP = TRUE: nao_correferente (46.0/18.0)
|  |  |  |  |  |  SP = FALSE: outra (581.0/269.0)
```

Árvore de Decisão do EXPERIMENTO 2 com atributos G1 e CF 0,35

```
TAM = TRUE: nao_correferente (271.0/70.0)
TAM = FALSE
|  NUM = TRUE: nao_correferente (6.0/1.0)
|  NUM = FALSE
|  |  APO_NP = TRUE: nao_correferente (33.0/8.0)
|  |  APO_NP = FALSE
|  |  |  SA = TRUE: nao_correferente (92.0/29.0)
|  |  |  SA = FALSE
|  |  |  |  NP_COM = TRUE: nao_correferente (55.0/19.0)
|  |  |  |  NP_COM = FALSE
|  |  |  |  |  PRE_ADJ = TRUE: nao_correferente (21.0/4.0)
|  |  |  |  |  PRE_ADJ = FALSE
|  |  |  |  |  |  SP = TRUE: nao_correferente (46.0/18.0)
|  |  |  |  |  |  SP = FALSE
|  |  |  |  |  |  |  PRE_NUM = TRUE: nao_correferente (12.0/5.0)
|  |  |  |  |  |  |  PRE_NUM = FALSE: outra (569.0/262.0)
```

Árvore de Decisão do EXPERIMENTO 2 com atributos G12 e CF 0,10

```
PRI_SENT = TRUE: nao_correferente (39.0)
PRI_SENT = FALSE
|  TAM = TRUE: nao_correferente (261.0/70.0)
|  TAM = FALSE
|  |  APO_NP = TRUE: nao_correferente (32.0/8.0)
|  |  APO_NP = FALSE
|  |  |  NUM = TRUE: nao_correferente (5.0/1.0)
|  |  |  NUM = FALSE
|  |  |  |  SA = TRUE: nao_correferente (89.0/29.0)
|  |  |  |  SA = FALSE
|  |  |  |  |  NP_COM = TRUE: nao_correferente (52.0/19.0)
|  |  |  |  |  NP_COM = FALSE
|  |  |  |  |  |  PRE_ADJ = TRUE: nao_correferente (21.0/4.0)
|  |  |  |  |  |  PRE_ADJ = FALSE
|  |  |  |  |  |  |  SP = TRUE: nao_correferente (46.0/18.0)
|  |  |  |  |  |  |  SP = FALSE: outra (560.0/248.0)
```

Árvore de Decisão do EXPERIMENTO 2 com atributos G12 e CF 0,35

```
PRI_SENT = TRUE: nao_correferente (39.0)
PRI_SENT = FALSE
|  TAM = TRUE: nao_correferente (261.0/70.0)
|  TAM = FALSE
|  |  APO_NP = TRUE: nao_correferente (32.0/8.0)
|  |  APO_NP = FALSE
|  |  |  NUM = TRUE: nao_correferente (5.0/1.0)
|  |  |  NUM = FALSE
|  |  |  |  SA = TRUE: nao_correferente (89.0/29.0)
|  |  |  |  SA = FALSE
|  |  |  |  |  NP_COM = TRUE: nao_correferente (52.0/19.0)
|  |  |  |  |  NP_COM = FALSE
|  |  |  |  |  |  PRE_ADJ = TRUE: nao_correferente (21.0/4.0)
|  |  |  |  |  |  PRE_ADJ = FALSE
|  |  |  |  |  |  |  SP = TRUE: nao_correferente (46.0/18.0)
|  |  |  |  |  |  |  SP = FALSE
|  |  |  |  |  |  |  |  COP = TRUE: nao_correferente (36.0/17.0)
|  |  |  |  |  |  |  |  COP = FALSE: outra (524.0/229.0)
```

Árvore de Decisão do EXPERIMENTO 2 com atributos G123 e CF 0,10

```
SEM_ANT = TRUE: nao_correferente (737.0/174.0)
SEM_ANT = FALSE
|  NUM = TRUE: nao_correferente (5.0)
|  NUM = FALSE
|  |  PRI_SENT = TRUE: nao_correferente (3.0)
|  |  PRI_SENT = FALSE
|  |  |  SUP = TRUE: nao_correferente (6.0/1.0)
|  |  |  SUP = FALSE: outra (354.0/68.0)
```

Árvore de Decisão do EXPERIMENTO 2 com atributos G123 e CF 0,35

```
SEM_ANT = TRUE: nao_correferente (737.0/174.0)
SEM_ANT = FALSE
|  NUM = TRUE: nao_correferente (5.0)
|  NUM = FALSE
|  |  PRI_SENT = TRUE: nao_correferente (3.0)
|  |  PRI_SENT = FALSE
|  |  |  SUP = TRUE: nao_correferente (6.0/1.0)
|  |  |  SUP = FALSE
|  |  |  |  SA = TRUE
|  |  |  |  |  COP = TRUE
|  |  |  |  |  |  SP = TRUE: outra (3.0/1.0)
|  |  |  |  |  |  SP = FALSE: nao_correferente (4.0/1.0)
|  |  |  |  |  COP = FALSE
|  |  |  |  |  |  TAM = TRUE: nao_correferente (4.0/1.0)
|  |  |  |  |  |  TAM = FALSE: outra (24.0/9.0)
|  |  |  |  SA = FALSE: outra (319.0/52.0)
```