Universidade de São Paulo – USP

Anotação Lingüística em XML do *Corpus* PLN-BR

Mírian Bruckschen, Fernando Muniz, José Guilherme C. de Souza,
Juliana Thiesen Fuchs, Kleber Infante, Marcelo Muniz,
Patrícia Nunes Gonçalves, Renata Vieira e Sandra Aluísio

NILC-TR-09-08
Junho 2008

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC – ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Este relatório apresenta uma proposta de esquema de anotação lingüística em XML que possa vir a ser adotado na construção e integração, no contexto do projeto PLN-BR, de *corpora* anotados da Língua Portuguesa. Para a construção da proposta, padrões internacionais e formatos de anotação utilizados por diversos grupos de pesquisa foram analisados. Apresentamos, ainda, o formato individual adotado por diferentes ferramentas de processamento de linguagem natural utilizadas em projetos envolvendo a Língua Portuguesa. As características de cada formato estudado são apresentadas e discutidas. A partir da identificação de princípios gerais propostos pelos padrões internacionais e dos pontos positivos e negativos dos modelos existentes, defendemos a proposta apresentada que busca integrar diferentes níveis de anotação lingüística em esquema XML.

Capítulo 1

Introdução

Para que a *Web* Semântica estabeleça-se em larga escala, faz-se necessário que um grande número de documentos seja anotado. Para esse propósito, ferramentas que possibilitam a anotação semi-automática de documentos têm sido desenvolvidas. No entanto, dado o grande número de documentos na *Web* que consistem total ou parcialmente de texto não estruturado, mostra-se necessário o desenvolvimento de ferramentas que possibilitem a análise automática da estrutura semântica de documentos textuais [4].

Assim, documentos de texto não estruturado serão transformados em semi-estruturados através da extração automática de partes importantes (extração de informações) e organizados através de agregamento ou classificação (mineração de texto). Para ambos objetivos acima, é preciso fazer a anotação lingüística dos textos.

Existem diversas ferramentas para anotação manual e automática de *corpus* com informações lingüísticas de vários níveis. Essas informações devem ser armazenadas de uma forma eficiente. Por eficiente, entende-se que os repositórios de dados com anotações lingüísticas devem permitir a expansão e a facilidade de uso e reuso dessas informações. Por se tratar de uma área recente, modelos de anotação que atendam às exigências citadas ainda estão sendo estudados.

Neste relatório, apresentamos a proposta de um padrão de anotação lingüística para uso da comunidade de processamento de linguagem natural da Língua Portuguesa, no contexto do projeto PLN-BR, baseado na linguagem de marcação XML, que atenda às características acima. Para isso, apresentamos uma revisão dos padrões de anotação para recursos lingüísticos que têm sido discutidos e desenvolvidos pelos grupos ISO TC37 SC 4 (International Standards Organization-Language Resources Standards), padrões utilizados por projetos, como o MuchMore e o TIGER. Falaremos também sobre formatos de anotação adotados por ferramentas que são utilizadas por trabalhos relacionados ao processamento da Língua Portuguesa, entre elas, o PALAVRAS, a MMAX e a RSTTool.

O restante deste documento está organizado da seguinte forma: o Capítulo 2 aborda o Projeto PLN-BR, ao qual a presente proposta encontra-se vinculada, e no Capítulo 3 são apresentadas ferramentas utilizadas em trabalhos de processamento de linguagem natural. No Capítulo 4, são apresentados o XCES e os padrões de anotação estudados pelos grupos TC37 SC 4 da ISO. Através da análise e crítica dos modelos existentes, queremos chegar a uma proposta integradora para atender anotação de *corpora* da Língua Portuguesa, considerando vários níveis lingüísticos. Com base no que foi apresentado nos capítulos anteriores, tal proposta de integração também é apresentada no Capítulo 4.

Capítulo 2

Projeto PLN-BR

O projeto Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil (PLN-BR), submetido ao CNPq no âmbito do edital CTInfo/MCT/CNPq nº 011/2005, e aprovado para o biênio 2006/2007, tem por objetivo geral a construção de um espaço interinstitucional de interação e intercâmbio de práticas de análise e investigação lingüístico-computacional acerca da representação e da recuperação de informação de natureza semântica e pragmático-discursiva veiculada por enunciados produzidos em português brasileiro. Subdividido em 7 subprojetos relativamente autônomos (Tabela 2.1), mas que compartilham o mesmo ponto de partida - qual seja, o tratamento da informação mobilizada em um mesmo *corpus* do português do Brasil. Este grande *corpus* do gênero informativo, subgênero jornalístico, chamado PLN-BR FULL gerou dois outros menores, chamados de PLN-BR CATEG e PLN-BR GOLD que serão descritos na Seção 2.1.

SubProjeto	**Responsável**	**Instituição**
Construção, Manutenção e Disponibilização de *corpora*	Sandra Maria Aluísio	USP/S.Carlos
Anotação de *corpora*	Renata Vieira	UNISINOS PUC/RS
Glosagem da Wordnet.Br e sua Indexação à WordNet de Princeton	Bento Carlos Dias-da-Silva	UNESP/Araraquara
Aprendizagem Automática de Informações Lexicais	Violeta de San Tiago Dantas Barbosa Quental	PUC/RJ
Sumarização Automática e Recuperação da Informação Textual	Lúcia Helena Machado Rino	UFSCar
Categorização de Textos	Vera Lúcia Strube de Lima	PUC/RS
Representação do Conhecimento Textual	Ronaldo Martins	MACKENZIE

Tabela 2.1: Edições Subprojetos do PLN-BR e seus responsáveis.

O projeto vincula pesquisadores vinculados à Universidade de São Paulo (USP), campus de São Carlos; à Universidade Federal de São Carlos (UFSCar); à Universidade Estadual Paulista (UNESP), campus de

Araraquara; à Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS); à Pontifícia Universidade Católica do Rio de Janeiro (PUCRJ); à Universidade do Vale do Rio dos Sinos (UNISINOS) e à Universidade Presbiteriana Mackenzie.

Cada um dos subprojetos possui problemas que lhe são específicos, mas são irmanados pelo mesmo ambiente interativo de desenvolvimento, pelo mesmo ponto de partida (o grande *corpus* e os dois outros derivados deste), pela construção colegiada de padrões e protocolos, e pela disposição para o reaproveitamento e a adaptação dos vários recursos já existentes desenvolvidos por vários dos proponentes. Um destes recursos, um **segmentador sentencial** do português do Brasil chamado SENTER [1], desenvolvido no NILC, foi adaptado para ser utilizado no projeto em tela. Outro, o **editor de cabeçalhos** desenvolvido no Projeto Lácio-*Web* [2], também foi adaptado para a edição do cabeçalho dos textos que seguem o padrão CES [3] adaptado posteriormente para XML, utilizado neste projeto. Este editor será incorporado na infra-estrutura de processamento de *corpus* sendo desenvolvida. Entretanto, como o foco deste relatório é a anotação dos textos, somente o cabeçalho e suas particularidades, por exemplo, a classificação em gênero e tipo textual utilizada, e não o editor em si, serão apresentados na Seção 2.2.

Sendo o *corpus* de trabalho o ponto de partida do projeto, a primeira questão de discussão foi o *corpus* do português do Brasil com que todos iriam trabalhar. Estes questionamentos e as decisões de amostragem, tamanho, gênero e domínio são apresentadas na Seção 2.1. A Seção 2.2 apresenta as decisões de projeto sobre a anotação básica dos textos dos córpus PLN-BR GOLD e PLN-BR CATEG, além das etapas de processamento dos textos recebidos da Folha de São Paulo até a sua disponibilização no Portal de *Corpus* [4].

2.1 *Corpus*

Os primeiros questionamentos sobre o *corpus* de trabalho do projeto PLN-BR foram levantados via fórum de discussão[5] e posteriormente via e-mail. As perguntas iniciais colocadas ao grupo foram:

1. O *corpus* envolveria um ou mais domínios?
2. Quais domínios?
3. Envolveria um ou mais tipos de texto?
4. Quais tipos de texto?
5. Quais seriam os critérios de compilação e balanceamento ideais?
6. Qual deveria ser o tamanho dessa base?
7. Em que formato deveria ser armazenada?
8. Como deveria ser disponibilizada?

E, posteriormente foram melhor elaboradas por um dos pesquisadores do grupo que questionou: a) se não devíamos usar a infra-estrutura de disponibilização e processamento do projeto Lácio-*Web*, já disponível no NILC; b) que talvez fosse interessante que não houvesse muita variação na forma e domínio dos textos do *corpus*; c) que um critério importante seria a relevância social do *corpus*; d) que o *corpus* devesse ser fechado, com textos mais simples, com estruturas sintáticas menos sofisticadas, que venham em prosa (e não em verso); e) que sejam reconhecidos por seu mérito informativo (e não pelo juízo estético); e que, f) de preferência, não sejam muito longos.

Falou-se em textos de patentes; dos domínios da Nanociência e Nanotecnologia e da bioinformática; na Bíblia; no Bulário Eletrônico da Anvisa[6]; no Guia de Remédios do UOL[7]; em textos relativos a um fato

[1] `http://www.icmc.usp.br/~taspardo/Senter.htm`
[2] `http://www.nilc.icmc.usp.br/lacioweb/`
[3] `http://www.cs.vassar.edu/CES/`
[4] `http://www.nilc.icmc.usp.br:8180/portal/`
[5] `http://www.ronaldomartins.pro.br/plnbr/forum/toc.htm`
[6] `http://www.anvisa.gov.br/e-bulas/`
[7] `http://www1.uol.com.br/remedios/`

histórico de importância para o Brasil; em textos de saúde pública, por exemplo, as cartilhas de órgãos governamentais em contraponto com textos científicos e de divulgação para leitores mais proficientes; em textos didáticos; e em textos da Wikipedia.

Entretanto, no primeiro Workshop do projeto realizado em São Carlos nos dias 16 e 17 de março de 2006 todos concordaram que o gênero de textos informativos, subgênero jornalístico era o que atenderia melhor a todos os subgrupos. Embora o NILC tivesse permissão de uso dos textos de 1994 da Folha de São Paulo (FSP), partimos para um pedido formal para a Folha, por ser o maior jornal do Brasil, em busca de dados mais atuais.

O primeiro pedido para a FSP foi de uma base de textos com a amostragem de ano sim ano não, meses fevereiro e agosto a partir de 1994, totalizando 12 meses. A partir desta grande base de textos tiraríamos uma amostra representativa para os trabalhos dos grupos. Uma grande e positiva surpresa foi a sugestão de um funcionário da FSP que trabalha com o banco de dados do jornal de utilizarmos uma amostragem que se pautasse pela idéia de "ano construído" que vem da idéia original de "semana construída". Nesta amostragem não privilegiaríamos um mesmo mês ou 2, no caso de termos as amostras de fevereiro e setembro, para não termos um *corpus* viciado, isto é, que supervalorizasse certos itens léxicos como é o caso de palavras relacionadas com o Carnaval, no caso do mês de fevereiro.

O ano construído para o projeto PLN-BR toma os textos de um mês aleatório de 1994 até um mês aleatório de 2005, totalizando 12 meses diferentes (Tabela 2.2).

Dia da Publicação do JORNAL	**Ano da Publicação do JORNAL**
1º a 31 de Janeiro	1994
1º a 31 de Maio	1995
1º a 30 de Junho	1996
1º a 31 de Outubro	1997
1º a 31 de Dezembro	1998
1º a 31 de Março	1999
1º a 31 de Julho	2000
1º a 30 de Setembro	2001
1º a 30 de Novembro	2002
1º a 30 de Abril	2003
1º a 28 de Fevereiro	2004
1º a 31 de Agosto	2005

Tabela 2.2: Edições autorizadas pela ESP para o Projeto PLN-BR.

As negociações com a FSP para obtenção da grande base de textos e de amostras representativas e balanceadas começaram em março de 2006 e em janeiro de 2007 o TERMO DE AUTORIZAÇÃO PARA UTILIZAÇÃO DE OBRA E OUTRAS AVENÇAS entre ICMC-USP (representando o Projeto PLN-BR) e a FSP foi assinado.

A grande base contém **125 mil** textos no formato Folio Views (um exemplo deste formato é apresentado na Seção 2.2). Entretanto, vários textos desta base eram compostos somente de informação de cabeçalho e estes não foram utilizados para fazerem parte da base de dados criada especialmente para o projeto PLN-BR.

Dadas as diferentes necessidades dos subgrupos quanto ao tamanho do *corpus* necessário, pedimos autorização de uso para 3 *corpus*, mantidos em 3 bases diferentes, mas com mesma estrutura:

1. PLN-BR FULL que contém **103.080 mil** textos da FSP e 29.014.089 tokens já foi disponibilizado para download em setembro de 2006, principalmente para os membros dos subprojetos *Glosagem da Wordnet.Br e sua Indexação à WordNet de Princeton* e *Aprendizagem Automática de Informações Lexicais*. Este *corpus* só pode ser acessado na *Web* com senha, com visualização de 30% de cada texto via concordâncias, por exemplo, devido à lei de direitos autorais. O *corpus* pode ser explorado totalmente pelos participantes do projeto para tarefas de criação de léxicos, por exemplo.

   Foi distribuído em codificação unicode sendo que os textos possuem as informações de título, subtítulo (quando existe), autores, tipo de texto, caderno, ano, número de palavras, *keywords* (quando existem),

seguido do texto cru. O título, subtítulo e autores não ganham etiquetas e assim colaboram para a contagem de freqüência quando usados no processador de *corpus* Unitex. As outras meta-informações (tipo de texto, caderno, ano, número de palavras e *keywords*) utilizam etiquetas Unitex, como mostra a Figura 2.1.

Os textos foram agrupados por diretórios indicando os anos (1994 a 2005) e os diretórios estão zipados com o software winrar. O arquivo tem 141MB compactado e 400MB descompactado.

Estes textos passaram por um novo crivo exigido pela FSP em dezembro de 2006 para dar acesso somente aos textos cujos créditos eram da FSP na montagem dos dois outros *corpus* que prevêem acesso a textos integrais. Este novo *corpus* possui **96.868 textos** e 26.425.483 tokens (mantemos este novo *corpus* em uma base de dados diferente, que chamaremos aqui de PLN-BR FULL 2).

2. PLN-BR CATEG que possui **30 mil** textos e 9.780.220 tokens. Também só pode ser acessado com senha pelos membros, mas o acesso aos textos é integral. Este *corpus* visa atender o subgrupo Categorização de Textos. Ele é uma amostra aleatória estratificada e proporcional à distribuição do *corpus* PLN-BR FULL com relação aos textos dos cadernos do jornal. Ele é formado por 30% dos textos do *corpus* PLN-BR FULL e possui somente notícias e reportagens para as quais a Folha de São Paulo possui direitos de republicação. Este *corpus* contém o *corpus* PLB-BR GOLD.

3. PLN-BR GOLD que possui **1024** textos e 338.441 tokens. Pode ser acessado livremente via *Web*. O tamanho deste *corpus*, que receberá atenção da maioria dos subgrupos, foi decidido para representar 1% do *corpus* PLN-BR FULL de forma a conservar, proporcionalmente, a distribuição deste *corpus* maior. Ele é uma amostra aleatória estratificada e proporcional à distribuição do *corpus* PLN-BR FULL com relação aos textos dos cadernos do jornal. Ele é formado por 1% dos textos do *corpus* PLN-BR FULL, e possui somente notícias e reportagens para as quais a Folha de São Paulo possui direitos de republicação. Este *corpus* está contido no *corpus* PLB-BR CATEG.

Para selecionarmos somente notícias e reportagens para a composição dos *corpus* PLN-BR GOLD e CATEG, classificamos os textos de forma automática usando um **classificador de tipos de textos** treinado com os 40 tipos de textos do Projeto Lácio-*Web* no *corpus* montado para o projeto de doutorado de Rachel Aires que foi defendido no ICMC-USP em 2005 sob orientação da Profa. Sandra Aluísio (mais informação sobre o classificador encontrado no site do projeto[8]).

2.2 Anotação XCES Básica para os *corpus* PLN-BR GOLD e PLN-BR CATEG

Em relação à anotação, são dois os níveis de representação das informações presentes em projetos de *corpus* atuais (Figura 2.2): a anotação estrutural, foco deste capítulo, e a anotação lingüística.

A anotação estrutural compreende a marcação de dados externos e internos dos textos. Como dados externos entendemos a documentação do *corpus* na forma de um cabeçalho que inclui dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, resumo do texto (se houver), tipologia textual (que será comentada na Seção 2.2.4) e informação sobre a distribuição do *corpus*. Como dados internos, temos a anotação de segmentação do texto cru que cuida da: a) marcação da estrutura geral - capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras, e b) marcação da estrutura de subparágrafos - elementos que são de interesse lingüístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e ênfase. No processo de codificação utiliza-se um elemento chamado cabeçalho (dados externos) e um chamado corpo (texto cru mais anotação de segmentação).

A codificação em XML usada no projeto PLN-BR segue as decisões do Projeto do *American National Corpus* (ANC)[9] por ser este contemporâneo ao nosso e por utilizar a codificação XML *Corpus* Encoding Standard (XCES) [9] para dados primários (texto cru) e anotações. Os textos dos *corpus* **PLN-BR GOLD**

[8] `http://www.nilc.icmc.usp.br/nilc/projects/linguarudo.html`
[9] `http://americannationalcorpus.org/`

```
Globo News dá 'furo' mundial

FRANCISCO MARTINS DA COSTA

{tipo de texto Notícia,.N}
{caderno TV FOLHA,.N}
{ano 1999,.N}
{número de palavras 125,.N}
{keywords [TELEVISÃO] [OSCAR, 1999] [GLOBONEWS],.N}

Na madrugada de domingo para segunda-feira, o "Em Cima da Hora", da
Glogo news, deu em primeira-mão que "O Resgate do Soldado Ryan", de
Steven Spielberg, ganhou o Oscar de melhor filme.

Foi uma notícia literalmente exclusiva, afinal o vencedor para todo
o resto da humanidade foi "Shakespeare Apaixonado". Parabéns Central
Globo de Jornalismo! É de "furos" como esse que o telespectador
gosta.

Mas não são exclusividade dos canais de notícia. O cantor Vinny, ao
analisar as chances de "Central do Brasil", na tarde de domingo na
MTV, ponderou que a concorrência era forte. "Ouvi dizer que 'La
Dolce Vita' é um ótimo filme", disse. Pena que "A Vida é Bela" em
italiano seja "La Vita È Bella".

(FRANCISCO MARTINS DA COSTA)
```

Figura 2.1: Exemplo de texto no *corpus* PLN-BR FULL.

```
1. Documentação (cabeçalho): dados bibliográficos, conjunto
de caracteres, informações sobre a codificação, etc.
2. Anotação de Segmentação
      2.1 Estrutura Geral: capítulos, parágrafos, títulos,
notas, elementos gráficos, etc.
      2.2 Estrutura de Subparágrafos: sentenças, citações,
palavras, abreviações, nomes, datas, etc.
3. Anotação Lingüística: informações lingüísticas sobre
segmentos como, por exemplo, a etiquetação morfossintática
e sintática.
```

Figura 2.2: Níveis de anotação em córpus modernos.

e PLN-BR CATEG são anotados em nível de parágrafos e sentenças, além da anotação de dados externos na forma de um cabeçalho, como mostrado nas próximas seções.

2.2.1 Codificação do Documento

Os *corpus* GOLD e CATEG usam o padrão proposto pelo XCES[10] para Anotações *Standoff*. Cada documento lógico nestes dois *corpus* é conceitualmente um documento XML único que obedece ao *schema* XCES xcesDoc.xsd. Fisicamente, os dados primários (arquivo com o texto cru) e suas anotações estão armazenados em múltiplos documentos XML que formam um grafo direcionado referenciando os dados primários (e potencialmente, regiões definidas sobre outras anotações também). Os nós dos grafos são virtuais, localizados entre cada caractere dos dados primários. As margens definidas sobre os nós do grafo são nomeadas com estruturas de *features* contendo informações de anotação associada com a região definida entre as margens.

Cada documento lógico definido no projeto PLN-BR para os *corpus* GOLD e CATEG consiste dos seguintes arquivos:

Nomedoarquivo.xces.xml	O arquivo de cabeçalho (header) no formato XCES que especifica o texto cru (conteúdo) e os arquivos de anotação.
Nomedoarquivo.txt	Documento com os dados primários (texto cru).
Nomedoarquivo-logical.xml	Marcação standoff da estrutura lógica do documento.
Nomedoarquivo-s.xml	Marcação standoff das fronteiras de sentença.

Além desses quatro arquivos disponibilizamos uma versão que intercala os dados primários com as anotações *standoff*:

Nomedoarquivo.xml	Versão que intercala os dados primários com as anotações *standoff*, chamada aqui de *merged*.

Usamos as seguintes codificações respectivas aos arquivos acima apresentados:

- Header (extensão.xces.xml) - UTF-8
- Content (extensão.txt) - Unicode
- Logical (extensão -logical.xml) - UTF-8
- Segment (extensão -s.xml) - UTF-8
- Merged (extensão.xml) - UTF-8

O formato de representação que separa os dados primários das anotações oferece considerável flexibilidade para o uso dos *corpus* PLN-BR, por exemplo:

- O texto primário pode ser usado sem anotações ou com anotações se necessário;
- O usuário pode escolher trabalhar com uma anotação em particular independente dos textos;
- O *corpus* pode conter anotações de diferentes tipos, ou várias versões de um único tipo de anotação (por exemplo, múltiplas marcações de etiquetadores morfossintáticos (*taggers*)) sem problemas de compatibilidade;
- O projeto pode distribuir anotações independentes do texto para *download*, porque as anotações possuem links para os dados originais (conteúdo), assim qualquer usário que já fez *download* do *corpus* pode porteriormente somente baixar as novas anotações.

[10] `http://www.xces.org/schema/2003`

2.2.2 Processo de Criação de um documento XML contendo texto e anotações

O formato *standoff* fornece flexibilidade para os criadores e usuários do nosso *corpus*, mas, em muitos casos, os usuários irão querer usar o *corpus* com anotações *in-line*. Para isso, disponibilizamos uma versão intercalada (*merged*) do texto com as duas anotações de segmentação (*logical* e *segment*) em um único documento XML. O ANC disponibiliza uma ferramenta que pode ser utilizada para realizar essa operação de criar uma versão do texto intercalada (*merged*) com as anotações e pode ser acessada na página das ferramentas ANC, chamada ANC Merge Tool[11]. Esta foi a ferramenta utilizada neste projeto.

2.2.3 Anotações *Standoff*

Os conjuntos de fronteiras em um grafo de anotação estão representados em um ou mais arquivos de anotações *standoff*. Cada arquivo de anotação *standoff* inclui uma série de anotações consistindo de uma ou mais *features*, representadas em XML com as tags <struct > and <feat >, respectivamente. Cada <struct> especifica uma fronteira (isto é, o limite dos dados primários) com os atributos *from* e *to* que referenciam os nós dentro do conjunto de nós dos dados primários. Por exemplo, a primeira frase do texto ESPORTE_1997_640.txt do *corpus* PLB-BR GOLD é:

Membros de torcidas uniformizadas...

O padrão proposto pelo XCES para Anotações *Standoff* assume um nó entre cada caractere, como pode ser observado na Figura 2.3.

```
|M|e|m|b|r|o|s|   |d|e|   |t|o|r|c|i|d|a|s|   |u|n|i|f|o|r|m|i|z|a|d|a|s|
                          1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4
```

Figura 2.3: Padrão proposto pelo XCES para Anotações *Standoff*.

O texto completo no arquivo ESPORTE_1997_640.txt é mostrado na Figura 2.4, juntamente com os aquivos de anotação *standoff logical* (Figura 2.5), *segment* (Figura 2.6) e *merged* (Figura 2.7).

2.2.4 Decisões sobre o Cabeçalho

O cabeçalho XCES segue as *Recomendações do Expert Advisory Group on Language Engineering Standards* (EAGLES) que seguem as especificações do *TEI Guidelines for Electronic Text Encoding and Interchange do the Text Encoding Initiative*[12].

Ele é dividido em 4 elementos principais, mostrados na Figura 2.8, com destaque para o elemento **encodingDesc**, na Figura 2.9, que descreve, dentre outros elementos, a tipologia textual utilizada no *corpus* (*classDecl*). No projeto PLN-BR decidimos utilizar a mesma classificação do Projeto Lácio-*Web*.

A taxonomia do Lácio-*Web* é quadripartida[13] (Gênero, Distribuição, Tipo Textual e Domínio). Fizemos o mapeamento mostrado no Apêndice A entre as categorias XCES e as instâncias da taxonomia do Lácio-*Web*.

Entretanto, não colocamos o elemento *classDecl* nos cabeçalhos de cada documento dos *corpus* GOLD e CATEG por este ser extenso e ser igual para todos os documentos, mas o disponibilizamos no site do projeto. Na parte de classificação do texto no elemento *profileDesc*, o item *catRef* refencia a nossa taxonomia como

[11] http://www.americannationalcorpus.org/tools/index.html
[12] http://www.tei-c.org/P4X/index.html
[13] http://www.nilc.icmc.usp.br/lacioweb/classificacoes.htm

Membros de torcidas uniformizadas do Corinthians emboscaram na madrugada de ontem o ônibus em que a delegação do clube viajava para São Paulo, após a derrota por 1 a 0 para o Santos, na Vila Belmiro, pelo Brasileiro.
No km 45, após o trecho de serra da rodovia dos Imigrantes (sentido São Paulo), torcedores com camisa da Gaviões atravessaram um ônibus em que viajavam na pista, transformando-o numa barricada.
Quando o ônibus dos jogadores chegou, a torcida investiu contra ele, armada com pedras, paus e galhos arrancados de árvores.
O ônibus estava sem proteção da Polícia Rodoviária, porque a diretoria não fez o pedido.
No ataque, os agressores xingavam os jogadores de mercenários e visavam especialmente Souza, Mirandinha e Donizete embora os dois últimos nem estivessem ali.
Orientando pelos seguranças do clube, os jogadores fecharam as cortinas e deitaram no corredor.
O ataque durou cerca de dez minutos e deixou dois feridos: o meia Rincón, que recebeu estilhaços de vidro na perna, e o motorista do ônibus, com corte no supercílio.
O vice de Futebol, José Mansur Farhat, e o diretor Jorge Neme não estiveram no cerco. Temendo represálias, deixaram o estádio antes do fim do jogo.
A Gaviões negou ter sido autora do ataque, mas seus diretores se contradizeram sobre o ocorrido.
Um diretor corintiano disse que viu diretores da Gaviões e o próprio presidente, Douglas Deúngaro, no ataque. Mas pediu para não ser identificado: “Sei do que eles são capazes de fazer”.
Deúngaro, por outro lado, disse que chegou ao local por acaso e tentou conter os torcedores, que, segundo ele, não eram da Gaviões.
Mas a tônica no clube ontem era tentar abafar o caso. Ainda não registrou queixa na polícia e é provável que nem o faça.
Os grupos de oposição política no clube criticaram o elo entre a atual administração e a Gaviões.
O ataque surge em hora crítica para o Corinthians e para a Gaviões. O time está em 20º lugar no Brasileiro e corre risco de rebaixamento. Já a Gaviões, proibida como todas as uniformizadas de frequentar estádios paulistas, negociava com a PM e o Ministério Público um modo de retornar.
LEIA mais sobre o ataque ao ônibus do Corinthians nas págs. 4-3 e 4-4.

Figura 2.4: Texto ESPORTE_1997_640.txt.

```
<?xml version="1.0" encoding="UTF-8" ?>
   <cesAna xmlns="http://www.xces.org/schema/2003" version="1.0.4">
   <struct type="cesDoc" from="0" to="2193">
           <feat name="version" value="1.0.4" />
           <feat name="id" value="ESPORTE_1997_640" />
           <feat name="xmlns:xsi" value="http://www.w3.org/2001/XMLSchema-instance" />
           <feat name="xmlns:xlink" value="http://www.w3.org/1999/xlink" />
           <feat name="xmlns" value="http://www.xces.org/schema/2003" />
   </struct>
   <struct type="text" from="0" to="2192" />
   <struct type="body" from="1" to="2191" />
   <struct type="div" from="2" to="2190">
            <feat name="type" value="materia" />
    </struct>
   <struct type="p" from="3" to="219">
            <feat name="id" value="p1" />
    </struct>
       …
   <struct type="p" from="2120" to="2189">
           <feat name="id" value="p15" />
  </struct>
  </cesAna>
```

Figura 2.5: Arquivo *Logical* dos dados primários ESPORTE_1997_640.txt.

```
<?xml version="1.0" encoding="UTF-8" ?>
   <cesAna xmlns="http://www.xces.org/schema/2003" version="1.0.4">
   <struct type="s" from="3" to="219">
       <feat name="id" value="p1s1" />
 </struct>
    <struct type="s" from="220" to="413">
        <feat name="id" value="p2s1" />
 </struct>
    <struct type="s" from="414" to="538">
         <feat name="id" value="p3s1" />
  </struct>
...
    <struct type="s" from="2180" to="2189">
     <feat name="id" value="p15s2" />
  </struct>
  </cesAna>
```

Figura 2.6: Arquivo de *segment* dos dados primários ESPORTE_1997_640.txt.

```
<?xml version="1.0" encoding="UTF-8" ?>
<cesDoc version="1.0.4" id="ESPORTE_1997_640" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
   instance" xmlns:xlink="http://www.w3.org/1999/xlink"
   xmlns="http://www.xces.org/schema/2003">
<text>
<body>
<div type="materia">
<p id="p1">
<s id="p1s1">Membros de torcidas uniformizadas do Corinthians emboscaram na madrugada de
   ontem o ônibus em que a delegação do clube viajava para São Paulo, após a derrota por 1
   a 0 para o Santos, na Vila Belmiro, pelo Brasileiro.</s>
   </p>
<p id="p2">
<s id="p2s1">No km 45, após o trecho de serra da rodovia dos Imigrantes (sentido São
   Paulo), torcedores com camisa da Gaviões atravessaram um ônibus em que viajavam na
   pista, transformando-o numa barricada.</s>
   </p>
...
<p id="p13">
<s id="p13s1">Os grupos de oposição política no clube criticaram o elo entre a atual
   administração e a Gaviões.</s>
   </p>
<p id="p14">
<s id="p14s1">O ataque surge em hora crítica para o Corinthians e para a Gaviões.</s>
<s id="p14s2">O time está em 20° lugar no Brasileiro e corre risco de rebaixamento.</s>
<s id="p14s3">Já a Gaviões, proibida como todas as uniformizadas de frequentar estádios
   paulistas, negociava com a PM e o Ministério Público um modo de retornar.</s>
   </p>
<p id="p15">
<s id="p15s1">LEIA mais sobre o ataque ao ônibus do Corinthians nas págs.</s>
<s id="p15s2">4-3 e 4-4</s>
   </p>
   </div>
   </body>
   </text>
   </cesDoc>
```

Figura 2.7: Arquivo de *Merged* dos dados primários ESPORTE_1997_640.txt.

<fileDesc>
Contém informações sobre o texto codificado (distribuição, fonte, etc.).
<encodingDesc>
Contém informações sobre a maneira como o texto foi codificado.
<profileDesc>
Contém informações sobre vários aspectos do texto (língua usada, classificação do texto segundo a sua tipologia, os participantes de um texto falado e sua situação, anotações, etc.).
<revisionDesc>
Resume o histórico de revisão (cabeçalho, segmentação e lingüística) de um documento.

Figura 2.8: Elementos principais do arquivo de cabeçalho XCES.

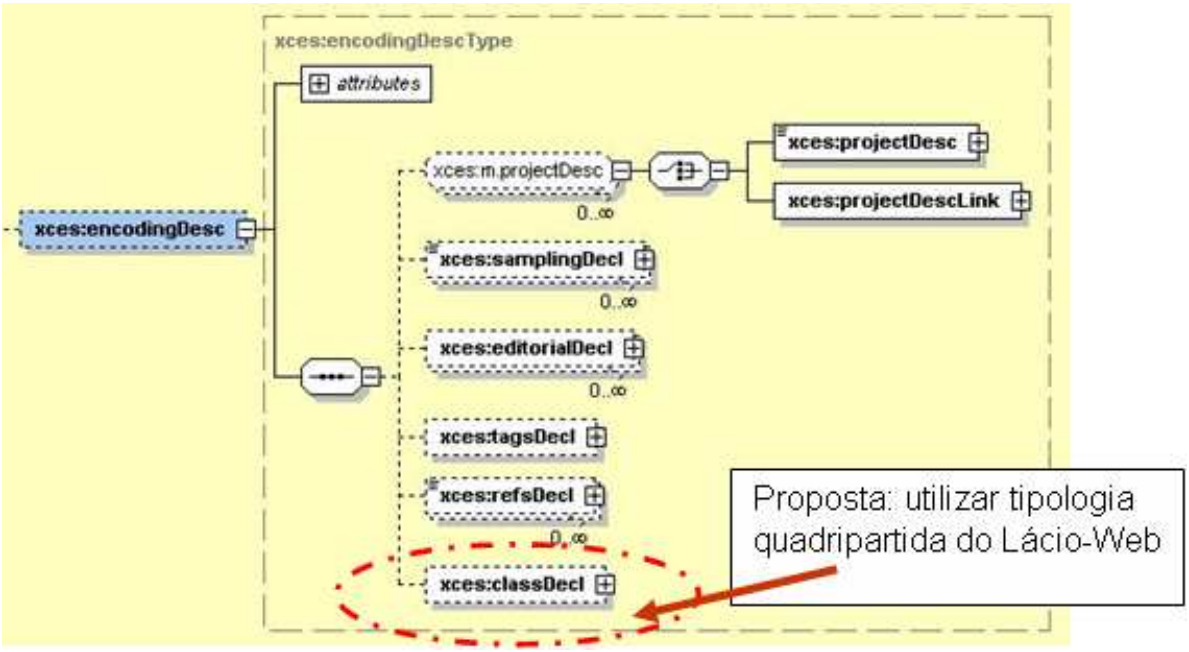

Figura 2.9: Esquema do elemento encodingDesc de um cabeçalho XCES.

descrita anteriormente, instanciando para o texto em foco. Por exemplo, para o texto mostrado na Figura 2.4, a classificação textual é a seguinte:

<catRef target="genero.8 genero.8.18 genero.8.18.10 distribuicao.12 tipotextual.35">

isto é, é um texto do gênero informativo *(genero.8.18)*, subgênero jornalístico *(genero.8.18.10)*, tipo textual reportagem *(tipotextual.35)*, distribuição jornal *(distribuicao.12)*. Os textos dos *corpus* GOLD e CATEG não foram anotados quanto ao domínio, assim não recebem os valores da taxonomia domínio também mostrada no Apêndice A.

2.2.5 Processamento do Header

Os arquivos que recebemos da FSP estavam no formato Folio Views. A Figura 2.11 traz um exemplo de cabeçalho e texto nesse formato. O primeiro passo foi separar as informações de cabeçalho do texto cru. Após isso, para selecionarmos somente notícias e reportagens, classificamos os textos de forma automática como citado na Seção 2.1.

Dos arquivos selecionados, processamos então seus cabeçalhos. Os seguintes campos que estavam presentes nos cabeçalhos foram utilizados no cabeçalho XCES: título, autores, responsável (crédito do texto), a data da publicação, o caderno, páginas e palavras chaves.

Utilizando essas informações mínimas de cabeçalhos, mais as informações do projeto PLN-BR (*projectDesc*) e de amostragem (*samplingDecl*), os textos foram inseridos de forma automática em nossa base de bados.

Em [1] apresentamos os itens do padrão XCES escolhidos para o cabeçalho do Projeto PLN-BR e o Editor *Web* de Cabeçalhos disponível no Portal de Córpus que abriga os três *corpus* do projeto. Com este editor a criação de um rico cabeçalho para os textos de um dado *corpus* a ser abrigado no Portal fica facilitada.

O trecho abaixo mostra o cabeçalho do texto apresentado na Figura 2.4.

```
<?xml version="1.0" encoding="UTF-8"?>
<cesHeader xmlns="http://www.xces.org/schema/2003" xmlns:xlink="http://www.w3.org/1999/xlink
    " xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.
    xces.org/schema/2003" version="1.0.4">
```

```
<fileDesc>
        <titleStmt>
                <title>1997out_1111</title>
                <respStmt>
                        <respType>Criação do Header</respType>
                        <respName type="person">Kleber Infante</respName>
                </respStmt>
                <respStmt>
                        <respType>Criação do Header</respType>
                        <respName type="person">Marcelo Muniz</respName>
                </respStmt>
        </titleStmt>
        <extent>
                <wordCount>421</wordCount>
                <byteCount units="bytes">5106.0</byteCount>
                <extNote>2</extNote>
        </extent>
        <sourceDesc>
                <biblStruct>
                        <monogr>
                                <title>Excel fica distante da crise corintiana</
                                    title>
                                <title>Patrocinador do time, responsável por
                                    investimento de US$ 26,6 milhões, posa de
                                    observador da situação</title>
                                <author>da Reportagem Local</author>
                                <respStmt>
                                        <respType>crédito</respType>
                                        <respName type="institution">DA REPORTAGEM
                                            LOCAL</respName>
                                </respStmt>
                                <imprint>
                                        <pubPlace>Folha de São Paulo</pubPlace>
                                        <publisher type="org">Empresa Folha da Manhã
                                             S.A.</publisher>
                                        <pubDate>29/10/97</pubDate>
                                        <pubAddress>São Paulo</pubAddress>
                                </imprint>
                                <biblNote>ESPORTE</biblNote>
                                <biblScope type="PP">3-11</biblScope>
                        </monogr>
                </biblStruct>
        </sourceDesc>
</fileDesc>
<encodingDesc>
        <projectDesc>O projeto Recursos e Ferramentas para a Recuperação de
            Informação em Bases Textuais em Português do Brasil (PLN-BR) - CNPq/
            CTInfo #550388/2005-2 - está subdividido em 7 subprojetos relativamente
            autônomos, mas que compartilham o mesmo ponto de partida - qual seja, o
            tratamento da informação mobilizada em um mesmo córpus do português do
            Brasil - e tem por objetivo geral a construção de um espaço
            interinstitucional de interação e intercâmbio de práticas de análise e
            investigação lingüístico-computacional acerca da representação e da
            recuperação de informação de natureza semântica e pragmático-discursiva
            veiculada por enunciados produzidos em português brasileiro. O projeto
            vincula pesquisadores da Universidade de São Paulo (USP), campus de São
            Carlos; da Universidade Federal de São Carlos (UFSCar); da Universidade
            Estadual Paulista (UNESP), campus de Araraquara; à Pontifícia
            Universidade Católica do Rio Grande do Sul (PUCRS); da Pontifícia
            Universidade Católica do Rio de Janeiro (PUCRJ); da Universidade do Vale
             do Rio dos Sinos (UNISINOS); e da Universidade Presbiteriana Mackenzie.
           </projectDesc>
        <samplingDecl>PLN-BR GOLD é o córpus gold standard do Projeto PLN-BR,
            formado por textos do jornal Folha de São Paulo que podem ser acessados
            integralmente na Web sem necessidade de senha de acesso. Ele é uma
            amostra aleatória estratificada e proporcional à distribuição do córpus
            global do projeto PLN-BR (chamado de PLN-BR FULL) com relação aos textos
```

```
                    dos cadernos do jornal. Ele é formado por 1% dos textos do córpus PLN-
                    BR FULL, o que equivale a 1.024 textos, e possui somente notícias e
                    reportagens para as quais a Folha de São Paulo possui direitos de
                    republicação. Este córpus está contido no córpus PLB-BR CATEG, também
                    criado no escopo do projeto PLN-BR. O córpus PLN-BR FULL, por sua vez, é
                    formado por 103,080 mil textos do jornal Folha de São Paulo, compondo
                    um ano construído a partir do ano de 1994 (toma um mês aleatório até o
                    ano de 2005). A classificação em notícias e reportagens foi feita de
                    forma automática usando-se um classificador de tipos de textos treinado
                    com os 40 tipos de textos do Projeto Lácio-Web (http://www.nilc.icmc.usp
                    .br/lacioweb/) no córpus montado para o projeto de doutorado de Rachel
                    Aires que foi defendido no ICMC-USP em 2005 sob orientação da Profa.
                    Sandra Aluísio (mais informação sobre o classificador
em http://www.nilc.icmc.usp.br/nilc/projects/linguarudo.html).</samplingDecl>
        </encodingDesc>
        <profileDesc>
                <textClass>
                        <catRef target="genero.8_genero.8.18_genero.8.18.10__distribuicao.12
                            _tipotextual.35_" />
                        <keywords>
                                <keyTerm><![CDATA[CRISE]]></keyTerm>
                                <keyTerm><![CDATA[BANCO]]></keyTerm>
                                <keyTerm><![CDATA[FUTEBOL]]></keyTerm>
                                <keyTerm><![CDATA[CORINTHIANS]]></keyTerm>
                                <keyTerm><![CDATA[CLUBE]]></keyTerm>
                                <keyTerm><![CDATA[EXCEL ECONÔMICO]]></keyTerm>
                        </keywords>
                </textClass>
                <annotations>
                        <annotation type="phrases" ann.loc="ESPORTE_1997_643-phrase.xml" >
                            phrases</annotation>
                        <annotation type="pos" ann.loc="ESPORTE_1997_643-pos.xml" >pos</
                            annotation>
                        <annotation type="tokens" ann.loc="ESPORTE_1997_643-token.xml" >
                            tokens</annotation>
                        <annotation type="logical" ann.loc="ESPORTE_1997_643-logical.xml"
                            >Logical markup</annotation>
                        <annotation type="s" ann.loc="ESPORTE_1997_643-s.xml" >Sentence
                            boundaries</annotation>
                        <annotation type="content" ann.loc="ESPORTE_1997_643.txt" >
                            Document content</annotation>
                </annotations>
        </profileDesc>
</cesHeader>
```

Dentro do nó *FileDesc*, no nó *titleStmt* o valor do *title* utilizado foi o nome do arquivo resultado do processamento dos arquivos que estavam no formato do Folio Views.

O valor do *respStmt* foi o mesmo em todos os textos inseridos automaticamente, especificando as pessoas que foram responsáveis pelo processamento e preenchimento do cabeçalho.

Dentro do *extent*, todos os campos foram preenchidos de forma automática, sendo que o valor do *extNote* representa o número de páginas do texto.

O *publicationStmt* possui o mesmo valor em todos os textos, nele colocamos o endereço do NILC como instituição que distribui o *corpus*.

Os campos dentro do *sourceDesc* foram preenchidos baseados nos campos encontrados no cabeçalho da FSP com exceção do *imprint* (informações da editora) que possui o mesmo valor para todos os textos.

No *biblNote* colocamos como valor o nome do caderno em que o texto foi publicado no jornal e no *biblScope* o número da página em que o texto se encontrava.

Dentro do nó *encodingDesc* somente foram preenchidos os campos *projectDesc* e *samplingDecl* e essas informações são compartilhadas entre todos os textos de um mesmo *corpus*.

Dentro do *profileDesc*, a classificação do texto (*textClass*) foi feita de forma automática e utilizamos a taxonomia apresentada no Apêndice A. No campo *keywords* utilizamos os valores das palavras-chaves dos textos encontrados no cabeçalho do Folio Views, quando presente.

```
<RD:REG><PS:Tit><FD:TITULO>Ajuda a mais pobres continua, diz Ruth<HR>
<PS:Data></FD:TITULO>10/03/1999<HR>
<PS:Ficha>Autor:  <FD:AUTOR></FD:AUTOR><PW:Popup,2,0.5>ODET160399;Fliv<LT><UN->
.<EL><PW:Popup,1,2.5>8035poli,brum ,TTAB ,<CR>
<LT><UN->  .<EL><TB><FD:IDENTIFICADOR>8035poli</FD:IDENTIFICADOR><CR>
Origem do texto: <FD:"COPYRIGHT_FOLHA">Da Redação</FD:"COPYRIGHT_FOLHA"><CR>
Editoria:       <FD:EDITORIA>BRASIL</FD:EDITORIA><TB>Página:       <FD:PAGINA>1-
6</FD:PAGINA><TB><FD:IDENTIFICADOR>3/3844</FD:IDENTIFICADOR><CR>
Edição:                                  <FD:CLICHE>Nacional</FD:CLICHE><TB>Tamanho:
<FD:TAMANHO>2516</FD:TAMANHO><CS:Num>    caracteres</CS><TB><FD:DATA>Mar    10,
1999</FD:DATA><CR>
Assuntos  Principais:  <FD:ASSUNTO>GOVERNO  FEDERAL;  CRISE  FINANCEIRA;  RUTH
CARDOSO; COMUNIDADE SOLIDÁRIA /PROGRAMA SOCIAL/</FD:ASSUNTO><HR>
<FD:TEXTO>Ajuda a mais pobres continua, diz Ruth <CR>
da Redação <CR>
Os cortes promovidos pelo governo federal em programas da área social não vão
prejudicar o atendimento à população mais carente, segundo a primeira-dama
Ruth Cardoso.<CR>
Em entrevista ao programa "Roda Viva", da TV Cultura, transmitido na última
segunda-feira, Ruth afirmou que quem está criticando a redução de recursos
para programas sociais (como os do Comunidade Solidária, do qual é presidente)
mistura "alhos com bugalhos".<CR>
Como  exemplo  disso,  citou  reportagem  publicada  pela  Folha  em  28  de
fevereiro.<CR>
...
Por outro lado, ao comentar o fato de que não há mulheres no ministério de seu
marido,  Ruth  afirmou  que  vê  dificuldade  no  fato  de  as  mulheres  "não
participarem tanto da política".<CR>
Questionada se poderia seguir o exemplo da primeira-dama dos EUA, Hillary
Clinton, e candidatar-se, no futuro, a um cargo eletivo, foi enfática: "Jamais
me candidatarei, não tenho vocação".<CR>
Por  fim,  ao  responder  a  um  apelo  de  um  telespectador  para  que  aconselhasse
FHC, disse: "Conselho sempre eu posso dar, mas isso não quer dizer que ele
siga".
```

Figura 2.10: Exemplo Cabeçalho no formato Folio Views.

Sobre as anotações, além do *content*, duas anotações foram criadas de forma automática: as anotações de *Logical markup* e *Sentence boundaries*.

Todas essas informações de cabeçalho, assim como os textos, foram inseridos de forma automática nos bancos de dados para os *corpus* GOLD e CATEG que compartilham a mesma estrutura.

O *corpus* GOLD possui além destas anotações citadas acima as anotações lingüísticas de POS, tokens e phrases, como mostrado no elemento profileDesc do cabeçalho do texto apresentado na Figura 2.4.

Após serem inseridos os textos e o cabeçalho mínimo, esses textos foram pós-processados. Primeiro foram processados pelo SENTER (versão em Java) e as anotações de segmentação foram geradas e inseridas nos bancos de dados também de forma automática. Além dos dados inseridos nos bancos de dados, uma cópia dos textos e também das anotações foram salvas em uma pasta especial. Um arquivo com a versão do texto em XML contendo as duas anotações e o próprio texto foi também gerada (versão *merged*). Para isso utilizamos bibliotecas criadas pelos pesquisadores que desenvolveram o ANC. O quinto arquivo gerado foi a versão em XML do cabeçalho.

Para padronizar os textos e as anotações de *Logical markup*, foram inseridos 2 espaços em branco e uma quebra de linha no começo e fim de todos os textos do *corpus*. Desta forma, os valores de *text*, *body* e *div* serão sempre 0, 1 e 2 e nunca vão se sobrepor. (Veja estes valores na Figura 2.5). A Figura 2.6 mostra um exemplo de marcação de *Sentence boundaries* que também foi gerada automaticamente pelo SENTER.

Capítulo 3

Ferramentas e formatos de codificação utilizados no processamento da língua portuguesa

Neste capítulo, apresentamos algumas ferramentas utilizadas em diferentes trabalhos relacionados ao processamento da língua portuguesa. As ferramentas apresentadas aqui são relacionadas ao processamento e anotação de *corpus* da língua portuguesa e se referem a diferentes níveis lingüísticos. Possuem, portanto, formatos de metadados lingüísticos independentes.

3.1 PALAVRAS

O analisador sintático PALAVRAS[1] [2] permite a anotação morfossintática de textos da Língua Portuguesa de forma automática. O analisador produz como saída o texto associado a informações morfológicas e lexicais de cada uma das palavras (tais como categorias gramaticais - substantivo, verbo, adjetivo, preposição - suas flexões de gênero e número, e, ainda, em alguns casos, seu tipo semântico). Além disso, identifica a formação de grupos de palavras com função específica numa sentença, como por exemplo, sujeito e predicado da sentença, ou sintagmas nominais e verbais. O PALAVRAS tem como saída três formatos: a primeira é uma forma gráfica, arbórea, da representação da análise do texto, como pode ser visto na Figura 3.1; a segunda, representa os mesmos dados, mas em formato texto, como pode ser visto na Figura 3.2, e a terceira, também representa os mesmos dados, mas em formato Tiger-XML (Seção 4.1.3), como pode ser visto na Figura 3.3.

Cada linha do formato texto representa a função sintática para aquele elemento ou grupo identificado. Por exemplo: S representa sujeito, P, predicado, H, núcleo e assim por diante. Depois dos dois pontos, a categoria ou forma sintática é dada para cada palavra ou grupo de palavras (np representa sintagmas nominais, n, substantivos, v, verbos). Entre parênteses está a forma canônica da palavra e as informações de flexão, gênero e número. Por último, a palavra é apresentada como no texto analisado. O símbolo = no início de cada linha representa o nível da expressão na árvore sintática. Uma descrição completa dos símbolos utilizados no PALAVRAS é dada em `http://visl.hum.sdu.dk/visl/pt/info/symbolset-manual.html`.

A representação gráfica produzida como saída pelo PALAVRAS não é facilmente interpretável para máquinas, mas é a mais adequada para humanos identificarem o significado da árvore sintática e dos seus componentes.

O formato texto disponibilizado pelo PALAVRAS, apesar de ser estruturado, ao contrário da representação gráfica, necessita que programas (*parsers*) específicos sejam escritos para sua interpretação. A codificação adotada não utiliza nenhum padrão e não segue princípios estudados pela ISO, como o de separabilidade e incrementabilidade. Além disso, sua legibilidade para humanos não é boa.

O formato Tiger-XML disponibilizado pelo PALAVRAS, assim como o formato texto, é estruturado e necessita de *parsers* específicos para ser interpretado. No entanto, diferentemente do formato texto, a

[1] `http://visl.hum.sdu.dk/visl/pt`

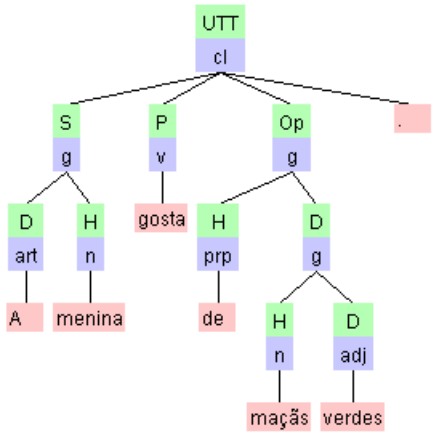

Figura 3.1: Codificação do texto “A menina gosta de maçãs verdes” no formato gráfico do PALAVRAS.

```
UTT:cl(fcl)
S:g(np)
=D:art('o' <artd> F S)  A
=H:n('menina' F S)      menina
P:v('gostar' fin  PR 3S IND)   gosta
Op:g(pp)
=H:prp('de')   de
=D:g(np)
==H:n('maçã' F P)maçãs
==D:adj('verde' F P)    verdes
```

Figura 3.2: Codificação do texto “A menina gosta de maçãs verdes” no formato VISL do PALAVRAS.

codificação adota um padrão de codificação das informações, também com legibilidade ruim para humanos, porém melhor que a do formato texto. O ponto negativo desta saída é que todas as informações sobre o texto (anotação) estão contidas em um único arquivo, utilizando o formato *in line* de anotação lingüística.

Considerando-se os três formatos de saída do PALAVRAS, o formato Tiger-XML pode ser considerado o mais adequado, visto que as informações contidas no arquivo de saída encontram-se em um nível estruturado, facilitando a extração de informações.

3.1.1 PALAVRAS XTRACTOR

O PALAVRAS XTRACTOR[8] é uma ferramenta para codificar em XML a saída do PALAVRAS (Seção 3.1). As informações produzidas pelo PALAVRAS XTRACTOR estão organizadas em três arquivos diferentes: *words*, *pos* e *chunks*.

O arquivo *words* armazena as palavras do texto. Consiste numa lista de elementos <word>, no qual cada valor é uma palavra do texto. Além disso, cada elemento <word> possui um identificador representado pelo atributo “id”. Um exemplo pode ser visto na Figura 3.4.

O arquivo *pos* contém as informações de *part-of-speech*, ou seja, as informações morfológicas de cada palavra, dentro do contexto do discurso. Nesse arquivo, as classes gramaticais das palavras e suas informações lexicais adicionais são sub-elementos dos elementos <word>. Na Figura 3.5, é possível ver um exemplo de uso desse arquivo.

No arquivo de informações lexicais, existe um elemento (linha 9 da Figura 3.5) que codifica as informações semânticas de uma palavra. Nessa linha, o valor do atributo `tag`, “Hfam”, é um valor semântico (definido pelo analisador PALAVRAS) que indica uma relação familiar ou similar. As informações semânticas são importantes e sua inclusão no esquema deve ser considerada.

Ainda no arquivo de informações lexicais, observamos que as classes gramaticais das palavras são representadas como elementos (linhas 3, 8 e 13 da Figura 3.5). Essa característica torna o processamento e o

```
<xml>
    <meta>
        <name>arboretum</name>
        <author>Eckhard Bick</author>
        <date>2004</date>
        <description>treebank</description>
        <format>VISL-TIGER</format>
    </meta>
<body>
<s id="s1" ref="1" source="Running text" forest="1" text="O Instituto Nacional de Pesquisas Espaciais
                                                          (Inpe) prediz um aumento de temperatura de até 5C
                                                          nas áreas mais secas da Amazônia, em 50 anos, se a
                                                          emissão de gases por queimadas permanecer nos níveis atuais.">
    <graph root="s1_500">
        <terminals>
            <t id="s1_1" word="O" lemma="o" pos="pron-indef" morph="DET M S" sem="--" extra="artd"/>
            <t id="s1_2" word="Instituto_Nacional_de_Pesquisas_Espaciais" lemma="Instituto_Nacional_de_Pesquisas_Espacia
            <t id="s1_4" word="(Inpe)" lemma="Inpe" pos="prop" morph="M/F S" sem="--" extra="np-close"/>
            <t id="s1_6" word="prediz" lemma="predizer" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="mv"/>
            ...
            <t id="s1_36" word="os" lemma="o" pos="pron-indef" morph="DET M P" sem="--" extra="-sam"/>
            <t id="s1_37" word="níveis" lemma="nível" pos="n" morph="M P" sem="Labs" extra="--"/>
            <t id="s1_38" word="atuais" lemma="atual" pos="adj" morph="M P" sem="--" extra="np-close"/>
            <t id="s1_39" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
        </terminals>
        <nonterminals>
            <nt id="s1_500" cat="s">
                <edge label="STA" idref="s1_501"/>
            </nt>
            <nt id="s1_501" cat="fcl">
                <edge label="S" idref="s1_502"/>
                <edge label="P" idref="s1_6"/>
                ..
            </nt>
</body>
</corpus>
</xml>
```

Figura 3.3: Codificação do texto “A menina gosta de maçãs verdes” no formato Tiger-XML do PALAVRAS.

```
<words>
 <word id="word_1">A</word>
 <word id="word_2">menina</word>
 <word id="word_3">gosta</word>
 <word id="word_4">de</word>
 <word id="word_5">maçãs</word>
 <word id="word_6">verdes</word>
 <word id="word_7">.</word>
</words>
```

Figura 3.4: Codificação do texto “A menina gosta de maçãs verdes” no formato PALAVRAS XTRACTOR.

```
1  <words>
2   <word id="word_1">
3    <art canon="o" gender="F" number="S">
4     <secondary_art tag="artd"/>
5    </art>
6   </word>
7   <word id="word_2">
8    <n canon="menina" gender="F" number="S">
9     <secondary_n tag="Hfam"/>
10   </n>
11  </word>
12  <word id="word_3">
13   <v canon="gostar">
14    <fin tense="PR" person="3S" mode="IND"/>
15   </v>
16  </word>
...
```

Figura 3.5: Codificação de *part-of-speech* do texto da Figura 3.4 no formato PALAVRAS XTRACTOR.

entendimento do arquivo difícil e confuso. O ideal seria que a classe gramatical da palavra fosse representada como um atributo, não como um elemento.

O terceiro e último arquivo é o arquivo *chunks*. Esse arquivo armazena a organização estrutural sintagmática representada pelos elementos <chunk>. Cada um desses elementos possui um identificador único, um `span` de palavras (que aponta para os arquivos *word* e *pos*) e outros dois atributos: `form` e `ext`. O atributo `form` nomeia o conjunto de palavras que desempenham o mesmo papel em uma sentença (como um sintagma nominal, por exemplo). O atributo `ext`, por sua vez, indica a função sintática desse conjunto (sujeito, predicado, entre outros). Um exemplo da codificação desse arquivo pode ser visto na Figura 3.6.

```
<text>
 <paragraph id="paragraph_1">
  <sentence id="sentence_1" span="word_1..word_7">
   <chunk id="chunk_1" ext="sta" form="fcl" span="word_1..word_6">
    <chunk id="chunk_2" ext="subj" form="np" span="word_1..word_2">
     <chunk id="chunk_3" ext="n" form="art" span="word_1"/>
     <chunk id="chunk_4" ext="h" form="n" span="word_2"/>
    </chunk>
    <chunk id="chunk_5" ext="p" form="v_fin" span="word_3"/>
    <chunk id="chunk_6" ext="piv" form="pp" span="word_4..word_6">
     <chunk id="chunk_7" ext="h" form="prp" span="word_4"/>
     <chunk id="chunk_8" ext="p" form="np" span="word_5..word_6">
      <chunk id="chunk_9" ext="h" form="n" span="word_5"/>
      <chunk id="chunk_10" ext="n" form="adj" span="word_6"/>
     </chunk>
    </chunk>
   </chunk>
  </sentence>
 </paragraph>
</text>
```

Figura 3.6: Codificação das informações estruturais do texto (sintagmas) no formato PALAVRAS XTRACTOR.

Uma característica positiva desse formato é a separação de informações de acordo com as diretrizes apontadas pelo XCES e pelo grupo ISO/TC 37/SC 4 (Capítulo 4). O texto é armazenado em um arquivo; as informações lexicais, em outro; e as informações estruturais, em um terceiro arquivo. Dessa forma, a informação fica melhor estruturada e pode ser facilmente expandida, acomodando outros níveis de informação lingüística (fonética, discursiva, semântica, entre outros). Além disso, este formato acomoda diversas informações no nível lexical que não aparecem em outros formatos (flexão de gênero e número, por exemplo).

3.2 MMAX

Expressões referenciais são expressões lingüísticas utilizadas para designar entidades ou objetos do mundo. Além disso, expressões referenciais podem ser utilizadas para introduzir entidades em um discurso ou podem fazer referência a entidades já mencionadas.

A anotação de referências associa expressões referenciais com informações que permitem a sua interpretação. Esse tipo de conhecimento é necessário para uma variedade de aplicações de processamento de linguagem natural, incluindo extração e busca de informações, tradução de máquina e diálogo homem-máquina [22].

Uma correferência é um tipo de expressão referencial. Ela consiste em duas ou mais expressões em um texto que se referem a uma mesma entidade do discurso. Quando uma entidade é referenciada pela primeira vez em um texto, a expressão que a descreve é dita como nova no discurso. Quando a entidade é retomada no texto, a expressão é dita como anafórica. A entidade sendo anafórica pode ser classificada ainda como direta, indireta ou associativa.

As expressões anafóricas diretas são aquelas antecedidas por uma expressão com o mesmo nome núcleo. Expressões anafóricas indiretas possuem o mesmo critério semântico de classificação e não possuem o mesmo nome núcleo do seu antecedente. As expressões anafóricas associativas introduzem um referente novo no discurso tendo alguma relação semântica com algum antecedente [21].

A ferramenta MMAX [17] é utilizada para a marcação manual de informações sobre correferência em discurso: textos e diálogos. A ferramenta possui uma interface gráfica (Figura 3.7) através da qual elementos de um texto podem ser selecionados pelo usuário e expressões referenciais como relações anafóricas e de correferência podem ser anotadas, gerando como saída, arquivos XML (Figura 3.8) contendo essas relações.

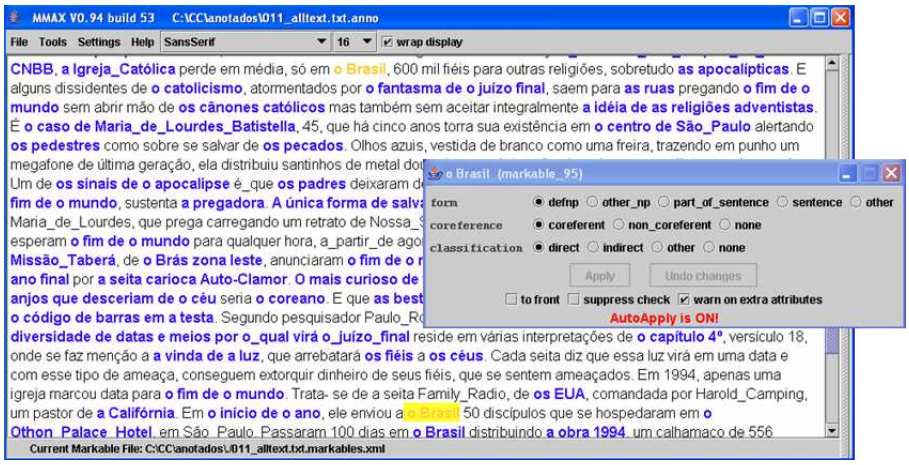

Figura 3.7: Exemplo da interface gráfica do MMAX. As palavras marcadas em azul representam sintagmas nominais. Os sintagmas marcados em amarelo possuem algum tipo de relação. Através dessa interface torna-se fácil a manipulação de sintagmas nominais e a definição das relações entre eles.

```
<markables>
 <markable id="markable_1" span="word_1..word_2" status="new"
      deitic="no" member="set_1" />
 <markable id="markable_2" span="word_7" is_anaphoric="indirect"
      status="old" pointer="markable_1" deitic="no" member="set_1" />
 ...
</markables>
```

Figura 3.8: Exemplo da saída XML do MMAX.

Para dar inicio à anotação manual, o MMAX utiliza como entrada de dados um arquivo XML que contém a codificação do texto (Figura 3.9). Cada elemento desse arquivo é uma palavra (*token*) do texto. Num primeiro momento, o anotador classifica e marca os sintagmas nominais e logo em seguida classifica se a expressão é correferente ou nova no discurso.

```
<words>
 <word id="word_1">A</word>
 <word id="word_2">menina</word>
 <word id="word_3">gosta</word>
 <word id="word_4">de</word>
 <word id="word_5">maçãs</word>
 <word id="word_6">verdes</word>
 <word id="word_7">.</word>
</words>
```

Figura 3.9: Arquivo de entrada do MMAX.

Como saída, a aplicação gera um arquivo XML chamado Markables (Figura 3.8) com os seguintes atributos: `id`, `span`, `pointer`, `member`, `np-form` e `status`.

O atributo `id` marca a identificação de cada marcação. O atributo `span` indica o início e o fim do sintagma nominal (sendo que os componentes desse intervalo são os elementos - palavras - que compõem o arquivo de *words* citado acima). Os atributos `pointer` e `member` codificam informações a respeito dos `markables`. O

primeiro aponta para o `id` do `markable` ao qual a expressão refere-se. Caso a expressão seja nova no discurso, esse atributo permanece vazio. O segundo, expressa relações indiretas entre diversos *markables* arbitrários e pode ser interpretado como uma operação de pertença de conjunto [18]. Por exemplo: *markables* com o mesmo valor no atributo `member` formam um conjunto de expressões que se referem a uma mesma entidade. O atributo `np-form` indica a classificação do sintagma nominal. O atributo `status` informa qual a classificação da expressão em relação a sua anaforicidade.

3.3 RSTTool

A *Rhetorical Structure Theory* (RST) é uma teoria que procura descrever a estrutura textual, ou seja, descrever as espécies de partes que formam o texto e os princípios de combinação das partes no texto inteiro. Considerando fundamental a noção de coerência textual, a RST atribui um papel a cada parte do texto. A teoria fornece um quadro para a investigação de proposições relacionais, que são proposições não determinadas, e sim inferidas, que surgem da estrutura do texto no processo de interpretação de textos.

As relações da estruturação do texto são funcionais, ou seja, o que importa é a categoria do efeito que elas produzem. Elas podem ser descritas em termos das finalidades do produtor textual, das suas suposições sobre o leitor e de determinados padrões proposicionais em relação ao conteúdo do texto. Segundo Mann et al. [15], as relações da estruturação do texto refletem as opções do produtor de organização e apresentação; é nesse sentido que a RST é retórica.

Os elementos-chave da RST são as relações e as extensões de texto (*text spans*), já que as definições de relações identificam determinados relacionamentos que podem acontecer entre duas extensões de texto. A noção de estrutura de um texto é definida em termos da rede de relações entre extensões de texto sucessivamente maiores. Conforme Mann e Thompson [16], dentro da estrutura relacional, a RST presume a homogeneidade, ou seja, haveria um grupo de padrões estruturais disponível para a organização do texto em cada escala da hierarquia. Esse grupo de padrões é identificado como esquemas da RST.

Em cada relação, há uma parte mais central, chamada núcleo, e uma parte mais periférica, chamada satélite; o núcleo e o satélite juntos formam a relação. Há também relações multinucleares, que se estabelecem não entre um núcleo e um satélite, e sim entre dois ou mais núcleos. Cada relação é definida segundo dois critérios: (1) condições (ou restrições), que inclui um grupo de condições para o núcleo e para o satélite individualmente e um grupo de condições para a combinação de núcleo e satélite (ou para a combinação de núcleos, no caso de relações multinucleares); (2) efeito, que inclui uma indicação do efeito que plausivelmente o produtor estava tentando produzir ao empregar a relação.

Cada critério de uma definição de relação especifica julgamentos particulares que o analista do texto deve fazer na construção da estrutura RST. O analista tem acesso ao texto, tem conhecimento do contexto no qual ele foi escrito e compartilha as convenções culturais do produtor textual e dos leitores pretendidos, mas não tem acesso direto nem ao produtor textual nem a outros leitores. Por isso, seus julgamentos sobre o produtor textual ou sobre os leitores devem ser mais de plausibilidade do que de certeza.

Para identificar as relações em um texto, o analista deve primeiramente dividí-lo em unidades, para depois identificar as extensões de texto e as relações entre elas. E, para determinar que relação acontece entre duas determinadas extensões de texto, o analista deve verificar se a definição da relação plausivelmente se aplica à unidade textual.

Cada unidade apresenta uma relação da RST. O tamanho da unidade é arbitrário para a RST, podendo abranger desde itens lexicais típicos até parágrafos inteiros, ou até unidades ainda maiores. Porém, as unidades devem ter integridade funcional independente. Mann, Matthiessen e Thompson [15], em geral, consideram orações como unidades; porém, os autores não apresentam critérios mais específicos para a segmentação de um texto em unidades. É possível encontrar esses critérios em autores que fazem uso da RST para desenvolver ferramentas de análise de textos. É o caso de Lynn Carlson e Daniel Marcu [5], que apresentam uma série de possibilidades de segmentação textual.

Segundo Carlson e Marcu [5], o primeiro passo para caracterizar a estrutura discursiva do texto é determinar as unidades discursivas elementares (*elementary discourse units*, EDUs), que seriam os mínimos blocos de construção de uma árvore discursiva. Assim, os autores escolheram a oração como a unidade elementar do discurso, usando indícios lexicais e sintáticos para ajudar na determinação de fronteiras. O passo seguinte em uma análise é estabelecer as relações que acontecem entre as unidades do texto. Conforme os autores,

uma vez determinadas as unidades elementares do discurso na segmentação do texto, extensões de texto adjacentes são ligadas a elas por meio de relações retóricas, o que cria uma estrutura hierárquica.

No site da RST[2], é possível encontrar informações sobre uma série de ferramentas que foram criadas para construir diagramas de análises RST. A ferramenta indicada para uso, no site, é a que foi desenvolvida por Mick O'Donnell, chamada RSTTool[3]. Essa ferramenta de análise RST permite visualizar as relações entre as unidades do texto e os esquemas de relações entre os diferentes níveis textuais.

Depois de inserido na ferramenta, o texto é segmentado em unidades, de acordo com os critérios do analista. O analista então relaciona essas unidades, podendo uní-las para estabelecer esquemas de relações em níveis progressivamente maiores, conforme ilustrado na Figura 3.10. As informações em XML referentes a Figura 3.10 são apresentadas na Figura 3.11.

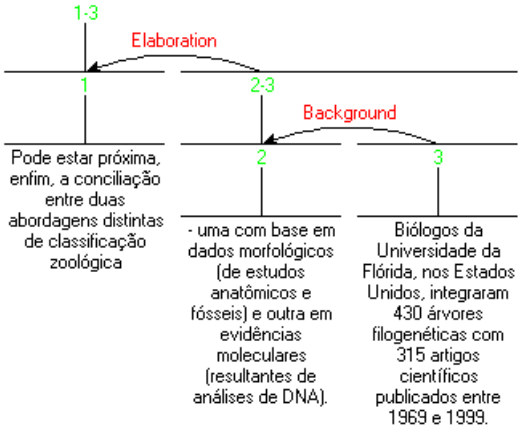

Figura 3.10: Tela de exemplo do RSTTool. Com essa interface torna-se mais fácil visualizar as unidades e definir as relações entre elas.

```
<rst>
 <header>
  <relations>
   <rel name="background" type="rst" />
   <rel name="elaboration" type="rst" />
   ...
  </relations>
 </header>
 <body>
  <segment id="1" parent="15" relname="span">
   Pode estar próxima, enfim ...
  </segment>
  <segment id="2" parent="16" relname="span">
   uma com base em dados ...
  </segment>

  <group id="17" type="span" parent="16"
   relname="background" />
  <group id="19" type="span" parent="18"
   relname="elaboration" />
  ...
  <group id="36" type="span" />
 </body>
</rst>
```

Figura 3.11: Arquivo de saída do RSTTool. Este arquivo contém as unidades e as relações entre elas.

O grupo ExtMT acrescenta um pequeno número de relações adicionais, cujas definições se encontram apenas no site da RST [14]. Há também o grupo de relações estabelecidas por Mick O'Donnell e o de relações

[2] www.sfu.ca/rst
[3] `http://www.wagsoft.com/RSTTool/`

estabelecidas por Daniel Marcu. Além desses grupos de relações já estabelecidas, existe a possibilidade, na ferramenta, de criar novas relações, redefinir as já existentes ou converter relações únicas em grupos de relações mais precisamente definidos. A ferramenta apresenta uma série de grupos de relações disponíveis para análises. O grupo de relações ClassicMT corresponde ao apresentado por Mann e Thompson [16].

A RST atualmente é estudada em diversas pesquisas, tanto estrangeiras quanto brasileiras. Entre elas, destacam-se as que têm como objetivo desenvolver ferramentas de análise e sumarização automática de textos.

Capítulo 4

Modelos de Anotação Lingüística

Nesse capítulo apresentamos alguns formatos de anotação lingüística utilizados por diversos projetos relacionados ao processamento de linguagem natural. Além disso, apresentamos nossa proposta de integração.

4.1 Padrões e Modelos de Anotação Lingüística

Alguns padrões internacionais para meta-dados lingüísticos estão em desenvolvimento. Nesta seção, apresentamos algumas idéias e princípios gerais propostos pelo grupo de padronização de codificação de *corpus* XCES e pelo grupo ISO TC37 SC4.

A ISO (*International Organization for Standardization*) formou um subcomitê (SC4) para questões referentes a recursos da linguagem. Os trabalhos desenvolvidos por esta organização têm como objetivo preparar padrões internacionais e *guidelines* para o uso efetivo de ferramentas que auxiliem no manejo de recursos lingüísticos. O trabalho tem como foco a modelagem de dados, marcação, troca de dados e avaliação de recursos da linguagem em detrimento de terminologias [12].

Dentro do SC4 foi criado o WG1, um grupo de trabalho que desenvolve um *framework* para a anotação lingüística que possa servir como uma referência ou pivô para diferentes esquemas de anotação e que possibilitará a sua integração e/ou comparação [12]. O grupo apontou os seguintes itens como requisitos para um *framework* para anotação lingüística:

- *Adequação expressiva*. É necessário oferecer meios para a representação de todas variedades das informações lingüísticas. Ou seja, fazer com que seja possível representar desde informações gerais até as informações mais específicas;
- *Independência de mídia*. O *framework* deve ser capaz de lidar com diversos tipos de mídia como imagem, texto, video, áudio, entre outros;
- *Adequação semântica*. As estruturas de representação devem ter uma semântica definida. Além disso, devem ser previstos meios centralizados de compartilhar descritores e categorias da informação;
- *Incrementabilidade*. Deve ser possível colocar os dados a qualquer momento, seja durante a anotação, seja somente na saída do processo de anotação;
- *Separabilidade*. Deve ser possível que uma aplicação consiga extrair ou separar facilmente informações específicas que ela esteja precisando para uma determinada tarefa;
- *Uniformidade*. Os dados devem ser representados utilizando uma nomenclatura padrão. Além disso, deve ser possível utilizar os mesmos métodos para combinar os dados;
- *Liberdade*. O *framework* não deve impor uma representação que dependa de uma única teoria lingüística;

- *Extensibilidade.* Deve ser possível declarar e trocar extensões com o registro central de categorias de dados;
- *Legibilidade.* As representações dos dados devem ser legíveis para humanos;
- *Consistência.* Diferentes mecanismos não devem ser usados para descrever o mesmo tipo de informação.

Além dos requisitos acima enumerados, o WG1 também apontou princípios de projeto para o desenvolvimento de um *framework* de anotação lingüística [12]. São eles:

- O modelo de dados e o documento anotado são distintos, mas são mapeáveis um para o outro;
- O modelo de dados é parcimonioso, genérico e formalmente preciso;
- O modelo de dados é construído numa clara separação entre estrutura e conteúdo;
- Existe um inventário de operações lógicas, suportadas pelo modelo de dados, que define a semântica abstrata do modelo;
- O documento anotado é altamente controlado pelo usuário;
- O mapeamento entre a forma flexível do documento anotado e o modelo de dados é feito via um formato de persistência arbitrário;
- O mapeamento do documento anotado para o formato final é documentado em um XML Schema (ou equivalente) associado ao documento;
- O mapeamento é operacionalizado tanto através de um processo de associação de dados ou através de um mapeamento que utiliza folhas de estilo (*schema-derived stylesheet*) entre o documento anotado e o formato final;
- Deve ser possível isolar camadas específicas da anotação de outras camadas (como, por exemplo, os dados originais). Isto é, deve ser possível que novas informações de anotação sejam adicionadas e somente apontem para os dados originais (*stand-off annotation*) ao invés destas estarem entre os dados;
- O formato final deve ser projetado para permitir a serialização e a desserialização de dados.

Estes requisitos são pontos interessantes de serem observados não só para o desenvolvimento de um *framework* para anotação lingüística, mas também para um formato de codificação de anotação lingüística. Tendo esses requisitos e princípios de projeto em mãos, é possível nortear o desenvolvimento de um formato para codificação de anotação lingüística que esteja em conformidade com padrões e práticas que vêm sendo estudados.

4.1.1 XCES

O XCES[1] é uma iniciativa para definição de padrões de anotação de *corpus* em XML[9]. É usado como *framework* de acesso e representação para o *American National Corpus*[9]. Esse padrão, entre outras coisas, provê especificações de codificação para a anotação linguística, junto com dados da arquitetura para o *corpora* linguístico [11][12].

O objetivo do XCES é especificar um formato que possibilite grande interoperabilidade entre anotações do mesmo fenômeno e tipos de anotação. O foco é propiciar um ambiente no qual anotações podem ser facilmente definidas e validadas, em vez de ditar o uso de valores e elementos específicos[10]. Em síntese, o XCES serve como uma interface entre diferentes tipos de anotações linguísticas. É possível usar as ferramentas desenvolvidas tanto para comparar as anotações de um texto escrito em línguas diferentes quanto para analisar duas anotações distintas do mesmo texto, escrito na mesma língua.

[1] `http://www.xces.org/`

A codificação do texto e de suas informações lingüísticas utilizando o XCES pode ser vista na Figura 4.1. Existe só um elemento definido, o elemento <struct>. Cada elemento desse tipo tem, obrigatoriamente, um atributo de identificação (`id`) e um atributo que informa o seu tipo (`type`). Na linha 1 da Figura 4.1, pode-se ver que o tipo do elemento é "S"(de *sentence*, frase, em inglês). O elemento da linha 8, tem como tipo o valor "NP"(*noun phrase*, sintagma nominal, em inglês). O XCES referencia as palavras do texto, utilizando a tecnologia XPath[6] através do atributo `xlink:href`. Esse atributo referencia o arquivo onde fica armazenado o texto.

```
1 <struct id="s0" type="S">
2    <struct id="s1" type="NP"
3         xlink:href="xptr(substring(/p/s[1]/text(), 1, 9))" rel="SBJ" />
4    <struct id="s2" type="VP"
5         xlink:href="xptr(substring(/p/s[1]/text(), 10, 14))" />
6    <struct id="s3" type="PP"
7         xlink:href="xptr(substring(/p/s[1]/text(), 16, 31))" rel="PIV">
8      <struct id="s4" type="NP"
9           xlink:href="xptr(substring(/p/s[1]/text(), 19, 31))" rel="P" />
10    </struct>
11 </struct>
```

Figura 4.1: Codificação do texto "A menina gosta de maçãs verdes" utilizando o XCES.

Nesse padrão, as informações obtidas no processo de anotação linguística não são adicionadas ao arquivo que contém as informações sobre o texto anotado. Ou seja, a informação da anotação não é mesclada com a original. Em vez disso, ela é armazenada em arquivos separados e relacionada ao original ou a outros documentos de anotação (*stand-off annotation*). O trabalho iniciado pelo XCES tem sido levado adiante pelo grupo ISO/TC 37/SC 4 [12].

4.1.2 MuchMore

O projeto MuchMore[2] foi desenvolvido por Paul Buitellar e Thierry Declerck no Laboratório de Inteligência Artificial da Alemanha[3]. O formato de anotação no projeto MuchMore integra múltiplos níveis da análise lingüística, organizando cada nível separadamente com opções de referência entre eles, utilizando índices (identificadores).

No formato de anotação do MuchMore, o texto é representado pelo elemento <text>. Esse elemento é composto por um ou mais elementos <token> que identificam as palavras do texto e também informações como referentes a palavra, por exemplo, a forma canônica e a informação de morfossintaxe. Para a análise das estruturas sintagmáticas, cada sintagma é delimitado pelo identificador dos elementos <token>. No exemplo da Figura 4.2, o primeiro sintagma nominal (*noun phrase* ou NP), é formado pelas palavras com os identificadores w1 e w2 ("A menina").

Nesse formato, as informações estruturais sintagmáticas do texto, bem como as morfossintáticas e o texto propriamente dito, estão armazenadas em um único arquivo. Essa característica pode ser negativa pois, caso sejam necessárias várias anotações de dados lingüísticos do mesmo texto, é preciso repetir o texto antes das informações lingüísticas em todos os arquivos. Isso causa redundância e desperdício de espaço. Outra característica observada foi a de que existem poucas informações mapeadas a respeito da morfossintaxe do texto (*part-of-speech*). Faltam informações como o gênero e o número da palavra, além da informação semântica. Como ponto positivo, destacamos o uso dos atributos `from` e `to` para especificar o intervalo dos sintagmas.

[2] `http://muchmore.dfki.de`
[3] `http://www.dfki.de/lt/`

```
<text>
 <token id="w1" pos="ART" lemma="o">A</token>
 <token id="w2" pos="NN" lemma="menina">menina</token>
 <token id="w3" pos="VB" lemma="gostar">gosta</token>
 <token id="w4" pos="PRP" lemma="de">de</token>
 <token id="w5" pos="NN" lemma="maçã">maçãs</token>
 <token id="w6" pos="ADJ" lemma="verde">verdes</token>
</text>
<chunks>
 <chunk id="c1" from="w1" to="w2" type="NP" head="w2">
 <chunk id="c2" from="w3" to="w3" type="VP">
 <chunk id="c3" from="w4" to="w6" type="PP" head="w4">
 <chunk id="c4" from="w5" to="w6" type="NP" head="w5">
</chunks>
```

Figura 4.2: Codificação do palavras e dos *chunks* do texto "A menina gosta de maçãs verdes" no formato utilizado pelo MuchMore.

4.1.3 Tiger-XML

O Tiger-XML[4] foi projetado para servir como um formato de interface. Esse formato é baseado em XML e equivalente à linguagem de descrição TIGER[13][25]. O Tiger-XML faz a interface de outros esquemas baseados em XML para a linguagem TIGER.

Nesse formato, o modelo de dados é baseado em grafos de sintaxe, isto é, grafos direcionados acíclicos com uma única raiz. Palavras, etiquetas de *part-of-speech*, etiquetas morfológicas e *lemmata* são atributos do elemento 'terminal'. Elementos não-terminais são representados através de um elemento chamado 'non-terminal' e referem-se ao terminal correspondente através de um identificador. Arestas secundárias são codificadas explicitamente [3].

Todos os nodos terminais e não-terminais são listados e as arestas são codificadas explicitamente como elementos. Um exemplo de uso desse formato de codificação pode ser visto na Figura 4.3.

```
1  <body>
2   <s id="s1">
3    <graph root="s1_502">
4     <terminals>
5      <t id="s1_1" word="A" pos="ART" morph="Def.Fem.S"/>
6      <t id="s1_2" word="menina" pos="NN" morph="Fem.S"/>
7      <t id="s1_3" word="gosta" pos="VTI"  morph=""/>
8      <t id="s1_4" word="de" pos="PREP" morph=""/>
9      <t id="s1_5" word="maçãs" pos="NN" morph="Fem.Pl"/>
10      <t id="s1_6" word="verdes" pos="ADJ" morph="Fem.Pl"/>
11    </terminals>
12    <nonterminals>
13     <nt id="s1_500" cat="NP">
14      <edge label="NK" idref="s1_1"/>
15      <edge label="NK" idref="s1_2"/>
16     </nt>
17     <nt id="s1_501" cat="PP">
18      <edge label="MO" idref="s1_4"/>
19      <edge label="HD" idref="s1_5"/>
20      <edge label="NK" idref="s1_6"/>
21     </nt>
22     <nt id="s1_502" cat="S">
23      <edge label="SB" idref="s1_500"/>
24      <edge label="HD" idref="s1_3"/>
25      <edge label="OI" idref="s1_501"/>
26     </nt>
27    </nonterminals>
28   </graph>
29  </s>
30 </body>
```

Figura 4.3: Codificação da frase "A menina gosta de maçãs verdes" no formato Tiger-XML.

O ponto negativo desse formato é o fato de a anotação morfossintática encontrar-se no mesmo arquivo

[4] `http://www.ims.uni-stuttgart.de/projekte/TIGER/\\TIGERSearch/doc/html/TigerXML.html`

que as informações a respeito do texto (palavras, sentenças e parágrafos) bem como as informações estruturais. Com isso, para gerar uma nova anotação linguística do texto, é necessário recodificar todo o texto novamente. Uma opção para que isso não aconteça seria separar as informações sobre o texto das informações morfossintáticas e estruturais.

Adicionalmente, apesar de o formato apresentar mapeamento para diversas informações morfológicas através do atributo `morph`, estas não estão representadas de forma estruturada. Na Figura 4.3, é possível ver que, na linha 5, a palavra "A"tem como valor a string "Def.Fem.S". Esse valor poderia ser dividido em vários atributos, estruturando os dados utilizando três atributos, um para cada tipo de informação morfológica da palavra.

A partir do Tiger-XML, pode-se facilmente montar graficamente a representação de uma árvore sintática. Isso se deve ao fato de o Tiger-XML representar os dados de maneira estruturada e organizada. Essa representação é feita através dos nodos terminais e não-terminais.

4.2 Proposta de Integração

Para o desenvolvimento da nossa proposta, levamos em consideração características presentes nos projetos apresentados na Seção 4.1, assim como os princípios estudados e desenvolvidos pela ISO. Além disso, a proposta tem como objetivo integrar os dados produzidos pelas ferramentas apresentadas na Seção 3.1. Como mostrado nessa seção, cada ferramenta possui um formato de anotação específico.

O MMAX, por exemplo, tem como entrada um arquivo XML no qual são codificadas as palavras do texto. Os resultados de outras ferramentas poderiam ser convertidos de modo a referenciar esse mesmo arquivo. Dessa forma, tanto as informações morfossintáticas, quanto as de correferência e relações retóricas do discurso poderiam ser associadas às palavras identificadas no arquivo de codificação do texto (que contém as palavras). A primeira proposta nesse sentido foi apresentada pelo PALAVRAS XTRACTOR (como visto na Seção 3.1.1).

Para a codificação dos arquivos optamos por utilizar o formato XCES. Essa decisão se deve ao fato de que esse formato é uma instância dos princípios discutidos no mesmo capítulo, desenvolvidos pelos maiores especialistas da área. Além disso, o formato é adotado por outros projetos de anotação de *corpora* como o ANC [9] e tem sido adotado cada vez mais. Adicionalmente, para a anotação de sentenças, paragráfos e cabeçalhos dos textos, o XCES já havia sido adotado. Lembramos que o objetivo deste trabalho é apresentar um formato padrão para a codificação de informações lingüísticas e não faria sentido utilizar formatos distintos para a anotação de diferentes níveis lingüísticos. Portanto, optamos por utilizar o XCES para a anotação de todo o *corpus* contemplando todos os seus níveis de informações.

Assim, dividimos as informações em vários arquivos, utilizando o princípio de projeto da separabilidade e mantivemos uma codificação básica para todos arquivos utilizando o princípio de uniformidade. Essa codificação básica é a mesma descrita na Seção 4.1. As únicas *tags* utilizadas são as *tags* <struct> e <feature>.

Os arquivos criados para armazenar as diferentes informações são: arquivo de codificação do texto (Figura 4.4), arquivo de informações lexicais (Figura 4.5), arquivo de informações estruturais sintagmáticas (Figura 4.6), arquivo de informações de relações anafóricas e correferenciais (Figura 4.7) e o arquivo de informações de relações retóricas (Figura 4.8).

Optamos por essa divisão pois, através dela, é possível desvincular o armazenamento do texto propriamente dito, das informações obtidas nos vários níveis de anotação lingüística. Além dos níveis lingüísticos aqui apresentados, outros poderiam ser facilmente acrescentados. Dessa forma, mantemos as informações organizadas e bem estruturadas conforme os princípios apresentados na Seção 4.1.

O arquivo mostrado no Figura 4.4 propõe um esquema de codificação para o texto. O atributo `type`, responsável pela identificação do tipo de anotação, recebe o valor *token* e os atributos `from` e `to` referenciam o arquivo base que contém o texto original. Este arquivo possui duas *feature*, uma responsável por identificar o token (`id`) e outra que contém a palavra base do texto (`base`).

O atributo `type` indica o tipo de estrutura descrita, que recebe o valor *pos*. As *features* que compõem esse nível de anotação são: identificador (`id`), classe gramatical (`class`), forma canônica (`canonical`), gênero (`gender`), número (`number`), tempo verbal (`tense`), forma verbal (`n_form`), modo verbal (`mode`), pessoa

```
<struct type="token" from="0" to="1">
    <feat name="id"      value="t1"/>
    <feat name="base"    value="A"/>
</struct>
<struct type="token" from="2" to="8">
    <feat name="id"      value="t2"/>
    <feat name="base"    value="menina"/>
</struct>
<struct type="token" from="9" to="14">
    <feat name="id"      value="t3"/>
    <feat name="base"    value="gosta"/>
</struct>
<struct type="token" from="15" to="17">
    <feat name="id"      value="t4"/>
    <feat name="base"    value="de"/>
</struct>
<struct type="token" from="18" to="23">
    <feat name="id"      value="t5"/>
    <feat name="base"    value="maças"/>
</struct>
<struct type="token" from="24" to="30">
    <feat name="id"      value="t6"/>
    <feat name="base"    value="verdes"/>
</struct>
<struct type="token" from="30" to="31">
    <feat name="id"      value="t7"/>
    <feat name="base"    value="."/>
</struct>
```

Figura 4.4: Codificação do texto "A menina gosta de maçãs verdes".

(`person`) e, ainda, uma referência para a palavra no texto (`token_ref`). De acordo com o valor do atributo `class`, diferentes *features* são utilizadas, como pode ser observado na Figura 4.5.

Caso o valor da *feature* `class` seja um valor que não indique um verbo, isto é, substantivo (valor "n"), adjetivo ("adj"), artigos ("art"), numerais ("num") e nomes próprios ("prop"), deverão estar presentes as *features* `gender` e `number`. Quando o valor de `class` indicar um verbo (valores "v-fin", "v-inf", "v-pcp2", "v-pcp1", etc.) as *features* obrigatórias são: `tense`, `n_form`, `mode` e `person`. Para todas as classes gramaticais, as *features* `id`, `canon` e `token-ref` são obrigatórias.

Realizando uma comparação desta abordagem e a utilizada pelo MuchMore (Seção 4.1.2) a principal mudança em relação ao formato, é a adoção da referência em relação ao token (`token_ref`). O formato proposto não repete a palavra no arquivo de informações morfossintáticas, apenas a referencia utilizando seu identificador. Além disso, outras *features* foram acrescentadas e utilizadas para especificar informações tais como informações semânticas (`semantic`) e refinamentos morfossintáticos (`complement`) a respeito da palavra no arquivo de informações morfossintáticas.

Para a concepção do arquivo com informações estruturais sintagmáticas do texto (Figura 4.6) utilizamos na etiqueta `type` o valor *phrase*. Neste nível de anotação é necessário informar a qual segmento de palavras a estrutura sintagmática se refere. Para isso, utilizamos os atributos `from` e `to` como apontadores para identificadores de *tokens* (no arquivo de *tokens*, Figura 4.4). As *features* que compõem esse nível de anotação são: identificador (`id`), categoria (`cat`), núcleo (`head`) e função sintática do sintagma (`function`).

O arquivo de codificação das relações anafóricas e correferenciais foi também estruturado para o padrão XCES. Este arquivo refere-se ao arquivo que codifica o texto (Figura 4.4) para determinar os *spans* dos *markables*. Essa estruturação das informações tomou-se como base o arquivo no formato original do MMAX (Figura 3.8). Em vez do atributo `span`, utilizamos dois atributos para especificar este intervalo: `from` e `to` assim como no arquivo de codificação dos dados estruturais sintagmáticos (Figura 4.6). Os valores destes, referem-se a palavras do arquivo de codificação do texto (que pode ser visto na Figura 4.4).

As demais *features* que compõem esse arquivo de anotação são: identificador único de *markable* (`id`), classificação dos sintagmas nominais (`np_form`), tipos de pronomes (`pro_form`), relações possíveis entre as entidades do discurso (`status`), tipo de relação associativa (`is_bridging`), identificador das cadeias de correferência (`member`), identificador de referência associativa (`pointer`), tipo de relação entre a entidade do discurso e o seu antecedente (`is_anaphoric`), como pode ser observado na Figura 4.7.

```
<struct type="pos">
    <feat name="id"       value="pos1"/>
    <feat name="class"    value="art"/>
    <feat name="gender"   value="F"/>
    <feat name="number"   value="S"/>
    <feat name="canon"    value="o"/>
    <feat name="complement" value="artd"/>
    <feat name="tokenref"  value="t1"/>
</struct>
<struct type="pos">
    <feat name="id"       value="pos2"/>
    <feat name="class"    value="noun"/>
    <feat name="gender"   value="F"/>
    <feat name="number"   value="S"/>
    <feat name="canon"    value="menina"/>
    <feat name="semantic"  value="Hfam"/>
    <feat name="tokenref"  value="t2"/>
</struct>
<struct type="pos">
    <feat name="id"       value="pos3"/>
    <feat name="class"    value="verb"/>
    <feat name="n_form"   value="fin"/>
    <feat name="tense"    value="PR"/>
    <feat name="person"   value="3S"/>
    <feat name="mode"     value="IND"/>
    <feat name="canon"    value="gostar"/>
    <feat name="tokenref"  value="t3"/>
</struct>
<struct type="pos">
    <feat name="id"       value="pos4"/>
    <feat name="class"    value="prp"/>
    <feat name="canon"    value="de"/>
    <feat name="tokenref"  value="t4"/>
</struct>
<struct type="pos">
    <feat name="id"       value="pos5"/>
    <feat name="class"    value="noun"/>
    <feat name="gender"   value="F"/>
    <feat name="number"   value="P"/>
    <feat name="canon"    value="maçã"/>
    <feat name="semantic" value="food-c"/>
    <feat name="tokenref"  value="t5"/>
</struct>
<struct type="pos">
    <feat name="id"       value="pos6"/>
    <feat name="class"    value="adj"/>
    <feat name="gender"   value="F"/>
    <feat name="number"   value="P"/>
    <feat name="canon"    value="verde"/>
    <feat name="tokenref"  value="t6"/>
</struct>
```

Figura 4.5: Codificação dos dados morfossintáticos do texto da Figura 4.4.

```
<struct type="phrase" from="t1" to="t2">
    <feat name="id"       value="phr1"/>
    <feat name="cat"      value="NP"/>
    <feat name="function" value="subj"/>
    <feat name="head"     value="t2"/>
</struct>
<struct type="phrase" from="t3" to="t3">
    <feat name="id"       value="phr2"/>
    <feat name="cat"      value="VP"/>
    <feat name="function" value="p"/>
    <feat name="head"     value="t3"/>
</struct>
<struct type="phrase" from="t4" to="t6">
    <feat name="id"       value="phr3"/>
    <feat name="cat"      value="PP"/>
    <feat name="function" value="piv"/>
    <feat name="head"     value="t4"/>
</struct>
<struct type="phrase" from="t5" to="t6">
    <feat name="id"       value="phr4"/>
    <feat name="cat"      value="NP"/>
    <feat name="function" value="p"/>
    <feat name="head"     value="t5"/>
</struct>
```

Figura 4.6: Codificação dos dados estruturais sintagmáticos do texto da Figura 4.4.

```
<struct type="markable"     from="t1" to="t2">
      <feat name="id"         value="mark1"/>
      <feat name="np_form"    value="def-np"/>
      <feat name="status"     value="new"/>
      <feat name="np_n"       value="yes"/>
</struct>
<struct type="markable"     from="t4" to="t7">
      <feat name="id"         value="mark2"/>
      <feat name="np_form"    value="indef-np"/>
      <feat name="np_n"       value="yes"/>
      <feat name="member"     value="set_1"/>
</struct>
<struct type="markable"     from="t9" to="t10">
      <feat name="id"         value="mark3"/>
      <feat name="np_form"    value="def-np"/>
      <feat name="status"     value="old"/>
      <feat name="np_n"       value="yes"/>
      <feat name="is_anaphoric" value="indirect"/>
      <feat name="member"     value="set_1"/>
</struct>
```

```
<struct type="markable"     from="t12" to="t14">
      <feat name="id"         value="mark4"/>
      <feat name="pro_form"   value="pes-pro"/>
      <feat name="np_n"       value="no"/>
      <feat name="member"     value="set_1"/>
</struct>
<struct type="markable"     from="t16" to="t17">
      <feat name="id"         value="mark5"/>
      <feat name="np_form"    value="def-np"/>
      <feat name="status"     value="associative"/>
      <feat name="np_n"       value="yes"/>
      <feat name="is_bridging" value="other-brigding"/>
      <feat name="pointer"    value="mark2"/>
</struct>
```

Figura 4.7: Codificação dos dados de relações anafóricas e correferenciais.

A codificação das informações retóricas deverá ser mapeada para o formato apresentado na Figura 4.8. A principal mudança em relação ao formato da ferramenta RSTTool (Figura 3.11) é a de que o texto anotado não é inserido no arquivo de informações retóricas, esse contém apenas referências para o arquivo que codifica o texto, compartilhado pelos outros níveis de anotação. Essas referências são feitas através dos atributos `from` e `to`.

Para seguir a padronização XCES as informações do arquivo da Figura 3.11, os elementos foram mapeados para as *features*: identificador (`id`), parent (`parent`), nome da relação (`relname`) e indicador de tipo núcleo de relação (`type`).

Ao final deste relatório, está anexado como apêndice, uma lista completa de todos os arquivos que compõem esta proposta e seus respectivos valores. São especificados, para cada arquivo, quais os valores possíveis do atributo `type` juntamente com suas *features* e seus conjuntos de valores possíveis.

4.3 Conclusão

A linguagem XML tem diversas vantagens que decorrem do fato de ser uma linguagem livre e flexível. Isso, no entanto, requer que um esforço maior seja necessário na definição de esquemas com propósitos específicos.

A proposta aqui apresentada levou em consideração diversas características identificadas como positivas em outros esquemas e evitou reproduzir as características negativas. Entre elas, a utilização de identificadores em todos os elementos, a separação das informações linguísticas distintas em arquivos diferentes e a estruturação da informação utilizando os atributos. Além disso, nossa proposta buscou acomodar as informações produzidas por diferentes ferramentas bastante difundidas de anotação, utilizada em diferentes trabalhos envolvendo o processamento da Língua Portuguesa, tais como o PALAVRAS, o MMAX e o RSTTool[7][24][23][20][26]. Como o modelo é extensível, outras ferramentas podem ser integradas ao modelo. Um exemplo é o DiZer [19].

```
<struct type="segment" from="t1" to="t10">
      <feat name="id"     value="seg1"/>
      <feat name="parent"  value="seg15"/>
      <feat name="relname" value="evaluation-s"/>
</struct>
<struct type="segment" from="t11" to="t17">
      <feat name="id"     value="seg2"/>
      <feat name="parent"  value="seg1"/>
      <feat name="relname" value="concession"/>
</struct>
<struct type="segment" from="t18" to="t26">
      <feat name="id"     value="seg3"/>
      <feat name="parent"  value="seg8"/>
      <feat name="relname" value="parenthetical"/>
</struct>
...
<struct type="group">
      <feat name="id"     value="grp40"/>
      <feat name="type"    value="mononuc"/>
      <feat name="parent"  value="seg1"/>
      <feat name="relname" value="purpose"/>
</struct>
<struct type="group">
      <feat name="id"     value="grp41"/>
      <feat name="type"    value="multinuc"/>
      <feat name="parent"  value="seg2"/>
      <feat name="relname" value="evidence"/>
</struct>
<struct type="group">
      <feat name="id"     value="grp42"/>
      <feat name="type"    value="mononuc"/>
      <feat name="parent"  value="grp40"/>
      <feat name="relname" value="same-unit"/>
</struct>
```

Figura 4.8: Codificação dos dados de relações retóricas.

Apêndice A

Taxonomia quadripartida em Gênero, Distribuição, Tipo Textual e Domínio do Projeto Lácio-*Web* (LW) e seu mapeamento no Elemento ClassDecl do cabeçalho XCES dos *corpora* do PLN-BR

O projeto LW distingue seus textos em quatro categorias ortogonais: gênero, tipo de texto, domínio e meio de distribuição.

O gênero discrimina o texto pela *intenção comunicativa* e pelo *caráter discursivo*, isto é, a comunidade (meio) em que circula e as atividades humanas que o tornam relevante. Convencionalizamos o uso de um super-gênero, chamado Literário (LT), um conjunto de gêneros e um conjunto de subgêneros. São 9 gêneros, apresentados na Tabela A.1, com seus respectivos subgêneros.

Gêneros	**Subgêneros**
Científico (CI)	- - - -
De referência (RE)	enciclopédico, lexicográfico, terminológico e outros
Informativo (IF)	jornalístico e outros
Jurídico (JU)	- - - -
Prosa (PR)[1†]	biografia, conto, novela, romance e outros
Poesia (PO)[†]	- - - -
Drama (DR)[†]	- - - -
Instrucional (IS)	didático, procedimental e outros
Técnico-Administrativo (TA)	- - - -

Tabela A.1: Gêneros e seus respectivos subgêneros.

O tipo textual trata-se de uma lista em constante atualização e que, no momento, é composta de 39 categorias (e "Outros" - tipos textuais não previstos). A seguir, essa relação, em ordem alfabética, é apresentada na Tabela A.2.

[1†]Esses gêneros, especialmente, advém do super-gênero Literário.

Apostila	Declaração	Manual	Parecer	Reportagem
Artigo	Decreto	Medida Provisória	Poema	Resenha
Ata	Edital	Memorando	Portaria	Resolução
Boletim	Editorial	Monografia	Projeto	Resumo
Carta	Ensaio	Notas Didáticas	Provimento	Sentença
Circular	Entrevista	Notícia	Receita	Súmula
Contrato	Lei	Ofício	Regimento	Testamento
Crônica	Livro-Texto	Outros	Relatório	Verbete

Tabela A.2: Atuais categorias do tipo textual.

Domínio é a "área de conhecimento" que tematiza a principal informação veiculada pelo texto. Temos 3 grandes linhas de domínio, denominadas "domínio geral". A cada uma dessas linhas associam-se subdomínios, denominados "domínios específicos". A divisão em termos de domínio geral apresenta as seguintes vertentes:

a) científica: classifica os textos tematizados pela ciência. Esse grupo é composto por 6 áreas do conhecimento abaixo demonstradas;

b) religião e pensamento: envolve os temas metafísicos, espirituais e teológicos (exemplo: livros de bruxaria, de auto-ajuda etc.);

c) generalidades: absorve os textos com temas variados e, de modo geral, inseridos num campo tematizado pelo senso comum (ex.: entretenimento). Inclui, além disso, os textos que abordam, de forma não-analítica, temas considerados pela ciência (exemplo: ciência e tecnologia, saúde, esporte etc.).

O meio de distribuição seleciona o canal através do qual o texto foi divulgado ao seu público-alvo. As possibilidades previstas pelo LW são as que estão apontadas na Tabela A.3, devidamente acompanhadas de suas siglas.

Cd-rom (CD)	Boletim/Informativo (BF)	Manuscrito (MA)
Código (CO)	Internet (IN)	Panfleto (PA)
Diário Oficial (DO)	Jornal (JO)	Periódico (PE)
Folheto (FO)	Livro (LI)	Revista (RE)
Tese (TE)	Dissertação (DI)	Não-identificado (NI)

Tabela A.3: Possíveis meios de distribuição previstos pelo LW.

Mostramos a seguir o elemento *ClassDecl* do cabeçalho XCES dos textos dos *corpus* do Projeto PLN-BR.

```
<?xml version="1.0" encoding="UTF-8"?>

<classDecl>
  <taxonomy id="genero">
    <category id="genero.7"><catDesc>Literário</catDesc>
      <category id="genero.7.1"><catDesc>Prosa</catDesc>
        <category id="genero.7.1.1"><catDesc>Biografia</catDesc></category>
        <category id="genero.7.1.2"><catDesc>Conto</catDesc></category>
        <category id="genero.7.1.3"><catDesc>Novela</catDesc></category>
        <category id="genero.7.1.4"><catDesc>Romance</catDesc></category>
      </category>
      <category id="genero.7.2"><catDesc>Poesia</catDesc></category>
      <category id="genero.7.3"><catDesc>Drama</catDesc></category>
      <category id="genero.7.4"><catDesc>Outros</catDesc></category>
```

```
    </category>
    <category id="genero.8"><catDesc>Não Literário</catDesc>
      <category id="genero.8.14"><catDesc>Técnico Administrativo</catDesc></category>
      <category id="genero.8.15"><catDesc>De Referência</catDesc>
        <category id="genero.8.15.6"><catDesc>Enciclopédico</catDesc></category>
        <category id="genero.8.15.13"><catDesc>Lexicografico</catDesc></category>
        <category id="genero.8.15.14"><catDesc>Terminologico</catDesc></category>
      </category>
      <category id="genero.8.16"><catDesc>Jurídico</catDesc></category>
      <category id="genero.8.17"><catDesc>Científico</catDesc></category>
      <category id="genero.8.18"><catDesc>Informativo</catDesc>
        <category id="genero.8.18.10"><catDesc>Jornalístico</catDesc></category>
      </category>
      <category id="genero.8.19"><catDesc>Instrucional</catDesc>
        <category id="genero.8.19.8"><catDesc>Didatico</catDesc></category>
        <category id="genero.8.19.9"><catDesc>Procedimental</catDesc></category>
      </category>
      <category id="genero.8.20"><catDesc>Outros</catDesc></category>
    </category>
  </taxonomy>
  <taxonomy id="distribuicao">
    <category id="distribuicao.1"><catDesc>Tese</catDesc></category>
    <category id="distribuicao.2"><catDesc>Dissertação</catDesc></category>
    <category id="distribuicao.3"><catDesc>Não-Identificado</catDesc></category>
    <category id="distribuicao.4"><catDesc>Revista</catDesc></category>
    <category id="distribuicao.5"><catDesc>Periódico</catDesc></category>
    <category id="distribuicao.6"><catDesc>Cd-rom</catDesc></category>
    <category id="distribuicao.7"><catDesc>Código</catDesc></category>
    <category id="distribuicao.8"><catDesc>Diário Oficial</catDesc></category>
    <category id="distribuicao.9"><catDesc>Folheto</catDesc></category>
    <category id="distribuicao.10"><catDesc>Boletim/Informativo</catDesc></category>
    <category id="distribuicao.11"><catDesc>Internet</catDesc></category>
    <category id="distribuicao.12"><catDesc>Jornal</catDesc></category>
    <category id="distribuicao.13"><catDesc>Livro</catDesc></category>
    <category id="distribuicao.14"><catDesc>Manuscrito</catDesc></category>
    <category id="distribuicao.15"><catDesc>Panfleto</catDesc></category>
</taxonomy>
  <taxonomy id="tipotextual">
    <category id="tipotextual.1"><catDesc>Artigo</catDesc></category>
    <category id="tipotextual.2"><catDesc>Apostila</catDesc></category>
    <category id="tipotextual.3"><catDesc>Boletim</catDesc></category>
    <category id="tipotextual.4"><catDesc>Circular</catDesc></category>
    <category id="tipotextual.5"><catDesc>Declaração</catDesc></category>
    <category id="tipotextual.6"><catDesc>Editorial</catDesc></category>
    <category id="tipotextual.7"><catDesc>Entrevista</catDesc></category>
    <category id="tipotextual.8"><catDesc>Lei</catDesc></category>
    <category id="tipotextual.9"><catDesc>Manual</catDesc></category>
    <category id="tipotextual.10"><catDesc>Verbete</catDesc></category>
    <category id="tipotextual.11"><catDesc>Livro-Texto</catDesc></category>
    <category id="tipotextual.12"><catDesc>Ensaio</catDesc></category>
    <category id="tipotextual.13"><catDesc>Edital</catDesc></category>
    <category id="tipotextual.14"><catDesc>Decreto</catDesc></category>
    <category id="tipotextual.15"><catDesc>Crônica</catDesc></category>
```

```
  <category id="tipotextual.16"><catDesc>Carta</catDesc></category>
  <category id="tipotextual.17"><catDesc>Ata</catDesc></category>
  <category id="tipotextual.18"><catDesc>Memorando</catDesc></category>
  <category id="tipotextual.19"><catDesc>Monografia</catDesc></category>
  <category id="tipotextual.20"><catDesc>Notícia</catDesc></category>
  <category id="tipotextual.21"><catDesc>Parecer</catDesc></category>
  <category id="tipotextual.22"><catDesc>Projeto</catDesc></category>
  <category id="tipotextual.23"><catDesc>Relatório</catDesc></category>
  <category id="tipotextual.24"><catDesc>Resenha</catDesc></category>
  <category id="tipotextual.25"><catDesc>Resumo</catDesc></category>
  <category id="tipotextual.26"><catDesc>Medida Provisória</catDesc></category>
  <category id="tipotextual.27"><catDesc>Notas Didáticas</catDesc></category>
  <category id="tipotextual.28"><catDesc>Ofício</catDesc></category>
  <category id="tipotextual.29"><catDesc>Portaria</catDesc></category>
  <category id="tipotextual.30"><catDesc>Receita</catDesc></category>
  <category id="tipotextual.31"><catDesc>Reportagem</catDesc></category>
  <category id="tipotextual.32"><catDesc>Resolução</catDesc></category>
  <category id="tipotextual.33"><catDesc>Regimento</catDesc></category>
  <category id="tipotextual.34"><catDesc>Outros</catDesc></category>
  <category id="tipotextual.35"><catDesc>Contrato</catDesc></category>
  <category id="tipotextual.36"><catDesc>Testamento</catDesc></category>
  <category id="tipotextual.37"><catDesc>Provimento</catDesc></category>
  <category id="tipotextual.38"><catDesc>Sentença</catDesc></category>
  <category id="tipotextual.39"><catDesc>Súmula</catDesc></category>
  <category id="tipotextual.40"><catDesc>Poema</catDesc></category>
</taxonomy>
<taxonomy id="dominio">
  <category id="dominio.2"><catDesc>Científico/Ciências Agrárias</catDesc>
    <category id="dominio.2.1"><catDesc>Agronomia</catDesc></category>
    <category id="dominio.2.2"><catDesc>Ciência e Tecnologia de Alimentos</catDesc>
    </category>
    <category id="dominio.2.3"><catDesc>Recursos Florestais e Engenharia
     Florestal</catDesc></category>
    <category id="dominio.2.4"><catDesc>Engenharia Agrícola</catDesc></category>
    <category id="dominio.2.5"><catDesc>Medicina Veterinária</catDesc></category>
    <category id="dominio.2.6"><catDesc>Recursos Pesqueiros e Engenharia de
      Pesca</catDesc></category>
    <category id="dominio.2.7"><catDesc>Zootecnia</catDesc></category>
  </category>
  <category id="dominio.3"><catDesc>Científico/Ciências Biológicas</catDesc>
    <category id="dominio.3.8"><catDesc>Biologia Geral</catDesc></category>
    <category id="dominio.3.9"><catDesc>Bioquímica</catDesc></category>
    <category id="dominio.3.10"><catDesc>Ecologia</catDesc></category>
    <category id="dominio.3.11"><catDesc>Genética</catDesc></category>
    <category id="dominio.3.12"><catDesc>Morfologia</catDesc></category>
    <category id="dominio.3.13"><catDesc>Parasitologia</catDesc></category>
    <category id="dominio.3.14"><catDesc>Biofísica</catDesc></category>
    <category id="dominio.3.15"><catDesc>Botânica</catDesc></category>
    <category id="dominio.3.16"><catDesc>Fisiologia</catDesc></category>
    <category id="dominio.3.17"><catDesc>Imunologia</catDesc></category>
    <category id="dominio.3.18"><catDesc>Microbiologia</catDesc></category>
    <category id="dominio.3.19"><catDesc>Zoologia</catDesc></category>
  </category>
```

```
 <category id="dominio.4"><catDesc>Científico/Ciências da Saúde</catDesc>
   <category id="dominio.4.20"><catDesc>Educação Física</catDesc></category>
   <category id="dominio.4.21"><catDesc>Enfermagem</catDesc></category>
   <category id="dominio.4.22"><catDesc>Farmácia</catDesc></category>
   <category id="dominio.4.23"><catDesc>Fisioterapia</catDesc></category>
   <category id="dominio.4.24"><catDesc>Fonoaudiologia</catDesc></category>
   <category id="dominio.4.25"><catDesc>Medicina</catDesc></category>
   <category id="dominio.4.26"><catDesc>Nutrição</catDesc></category>
   <category id="dominio.4.27"><catDesc>Odontologia</catDesc></category>
   <category id="dominio.4.28"><catDesc>Saúde Coletiva</catDesc></category>
   <category id="dominio.4.29"><catDesc>Terapia Ocupacional</catDesc></category>
 </category>
 <category id="dominio.5"><catDesc>Científico/Ciências Exatas e da Terra</catDesc>
   <category id="dominio.5.30"><catDesc>Geociências</catDesc></category>
   <category id="dominio.5.31"><catDesc>Engenharia Civil</catDesc></category>
   <category id="dominio.5.32"><catDesc>Matemática</catDesc></category>
   <category id="dominio.5.33"><catDesc>Engenharia de Minas</catDesc></category>
   <category id="dominio.5.34"><catDesc>Engenharia de Materiais</catDesc></category>
   <category id="dominio.5.35"><catDesc>Engenharia Elétrica</catDesc></category>
   <category id="dominio.5.36"><catDesc>Engenharia Mecânica</catDesc></category>
   <category id="dominio.5.37"><catDesc>Engenharia Química</catDesc></category>
   <category id="dominio.5.38"><catDesc>Engenharia Aeroespacial</catDesc></category>
   <category id="dominio.5.39"><catDesc>Outras Engenharias</catDesc></category>
   <category id="dominio.5.40"><catDesc>Engenharia Biomédica</catDesc></category>
   <category id="dominio.5.41"><catDesc>Engenharia Física</catDesc></category>
   <category id="dominio.5.42"><catDesc>Engenharia Sanitária</catDesc></category>
   <category id="dominio.5.43"><catDesc>Engenharia de Produção</catDesc></category>
   <category id="dominio.5.44"><catDesc>Engenharia Nuclear</catDesc></category>
  <category id="dominio.5.45"><catDesc>Engenharia de Transportes</catDesc></category>
<category id="dominio.5.46"><catDesc>Engenharia Naval e Oceânica</catDesc></category>
   <category id="dominio.5.47"><catDesc>Química</catDesc></category>
   <category id="dominio.5.48"><catDesc>Ciência da Computação</catDesc></category>
   <category id="dominio.5.49"><catDesc>Estatística</catDesc></category>
   <category id="dominio.5.50"><catDesc>Astronomia</catDesc></category>
   <category id="dominio.5.51"><catDesc>Física</catDesc></category>
   <category id="dominio.5.52"><catDesc>Mecânica</catDesc></category>
   <category id="dominio.5.53"><catDesc>Oceanografia</catDesc></category>
 </category>
   <category id="dominio.6"><catDesc>Científico/Ciências Humanas</catDesc>
   <category id="dominio.6.54"><catDesc>Antropologia</catDesc></category>
   <category id="dominio.6.55"><catDesc>Turismo</catDesc></category>
   <category id="dominio.6.56"><catDesc>Ciências Contábeis</catDesc></category>
   <category id="dominio.6.57"><catDesc>Filosofia</catDesc></category>
   <category id="dominio.6.58"><catDesc>História</catDesc></category>
<category id="dominio.6.59"><catDesc>Lingüística, Letras e Artes</catDesc></category>
   <category id="dominio.6.60"><catDesc>Psicologia</catDesc></category>
   <category id="dominio.6.61"><catDesc>Relações Públicas</catDesc></category>
   <category id="dominio.6.62"><catDesc>Arqueologia</catDesc></category>
   <category id="dominio.6.63"><catDesc>Ciência Política</catDesc></category>
   <category id="dominio.6.64"><catDesc>Geografia</catDesc></category>
   <category id="dominio.6.65"><catDesc>Jornalismo</catDesc></category>
   <category id="dominio.6.66"><catDesc>Pedagogia</catDesc></category>
   <category id="dominio.6.67"><catDesc>Publicidade e Propaganda</catDesc></category>
```

```
      <category id="dominio.6.68"><catDesc>Sociologia</catDesc></category>
      <category id="dominio.6.69"><catDesc>Teologia</catDesc></category>
    </category>
  <category id="dominio.7"><catDesc>Científico/Ciências Sociais Aplicadas</catDesc>
     <category id="dominio.7.70"><catDesc>Administração</catDesc></category>
      <category id="dominio.7.71"><catDesc>Arquitetura e Urbanismo</catDesc></category>
      <category id="dominio.7.72"><catDesc>Ciência da Informação</catDesc></category>
      <category id="dominio.7.73"><catDesc>Comunicação</catDesc></category>
      <category id="dominio.7.74"><catDesc>Desenho Industrial</catDesc></category>
      <category id="dominio.7.75"><catDesc>Direito</catDesc></category>
      <category id="dominio.7.76"><catDesc>Economia</catDesc></category>
      <category id="dominio.7.77"><catDesc>Serviço Social</catDesc></category>
    </category>
  <category id="dominio.8"><catDesc>Religião e Pensamento</catDesc>
      <category id="dominio.8.78"><catDesc>Auto-ajuda</catDesc></category>
      <category id="dominio.8.79"><catDesc>Magia e Bruxaria</catDesc></category>
      <category id="dominio.8.80"><catDesc>Religião</catDesc></category>
  </category>
 <category id="dominio.9"><catDesc>Generalidades</catDesc>
      <category id="dominio.9.81"><catDesc>Ambiente</catDesc></category>
      <category id="dominio.9.82"><catDesc>Celebrações/Solenidades</catDesc></category>
      <category id="dominio.9.83"><catDesc>Culinária</catDesc></category>
      <category id="dominio.9.84"><catDesc>Empreendimento</catDesc></category>
      <category id="dominio.9.85"><catDesc>Esporte</catDesc></category>
      <category id="dominio.9.86"><catDesc>Meio Ambiente</catDesc></category>
      <category id="dominio.9.87"><catDesc>Policial</catDesc></category>
      <category id="dominio.9.88"><catDesc>Dinheiro/Finanças</catDesc></category>
      <category id="dominio.9.89"><catDesc>Cotidiano/Comunidade</catDesc></category>
      <category id="dominio.9.90"><catDesc>Cultura</catDesc></category>
      <category id="dominio.9.91"><catDesc>Entretenimento, Lazer,
       Recreação</catDesc></category>
      <category id="dominio.9.92"><catDesc>Informática e Internet</catDesc></category>
      <category id="dominio.9.93"><catDesc>Moda</catDesc></category>
      <category id="dominio.9.94"><catDesc>Saúde</catDesc></category>
      <category id="dominio.9.95"><catDesc>Sociedade</catDesc></category>
      <category id="dominio.9.96"><catDesc>Viagem</catDesc></category>
      <category id="dominio.9.97"><catDesc>Educação</catDesc></category>
      <category id="dominio.9.98"><catDesc>Ciência e Tecnologia</catDesc></category>
      <category id="dominio.9.99"><catDesc>Justiça</catDesc></category>
      <category id="dominio.9.100"><catDesc>Ambiente Doméstico</catDesc></category>
    </category>
<category id="dominio.10"><catDesc>Outros</catDesc></category>
    </category>
  </taxonomy>
</classDecl>
```

Referências Bibliográficas

[1] S. M. ALUÍSIO, F. A. M. MUNIZ, and K. INFANTE. Projeto PLN-BR: o Cabeçalho em XML para os Textos do Córpus e o Editor Web de Cabeçalhos. Série de Relatórios do NILC (NILC-TR-07-05). São Carlos - SP, Junho 2007, 69 p. Technical report, 2007.

[2] E. Bick. *The Parsing System "PALAVRAS- Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Department of Linguistics, University of Århus, DK., 2000.

[3] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. The tiger treebank. In *Workshop on Treebanks and Linguistic Theories*, pages 24–42, Sozopol, 2002.

[4] P. Buitelaar and T. Declerck. *Annotation for the Semantic Web*, volume 96 of *Frontiers in Artificial Intelligence and Applications Series*, chapter Linguistic Annotation for the Semantic Web. IOS Press, 2003.

[5] L. Carlson and D. Marcu. Discourse tagging reference manual. Technical Report ISI-TR-545, ISI, 2001.

[6] J. Clark and S. DeRose. Xml path language (xpath). http://www.w3.org/TR/xpath, November 1999.

[7] C. F. da Silva, R. Vieira, and F. S. Osorio. Evaluating the use of linguistic information in the preprocessing phase of text mining. *Iberoamerican Journal of Artificial Intelligence*, 9(26):59–66, 2005.

[8] C. Gasperin, R. Vieira, R. Goulart, and P. Quaresma. Extracting xml syntactic chunks from portuguese corpora. In *Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages*, 2003.

[9] N. Ide, P. Bonhomme, and L. Romary. Xces: An xml-based encoding standard for linguistic corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*, 2000.

[10] N. Ide and L. Romary. Encoding syntactic annotation. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 281–296, Kluwer, Dordrecht, 2003.

[11] N. Ide and L. Romary. Outline of the international standard linguistic annotation framework. In *Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right*, pages 1–5, Sapporo, 2003.

[12] N. Ide and L. Romary. International standard for a linguistic annotation framework. *Journal of Natural Language Engineering*, 10:3-4:211–225, 2004.

[13] E. König and W. Lezius. The TIGER language - a description language for syntax graphs, Formal definition. Technical report, 2003.

[14] W. Mann. Relation definitions. Ultimo acesso em 04/2006.

[15] W. C. Mann, C. M. I. M. Matthiessen, and S. A. Thompson. *Discourse description: diverse linguistic analyses of a fund-raising text*, chapter Rhetorical Structure Theory and Text Analysis, pages 39–78. John Benjamins, Amsterdam, 1992.

[16] W. C. Mann and S. A. Thompson. Rhethorical structure theory: toward a functional theory of text organization. *Text*, 3(8):243–281, 1988.

[17] C. Muller and M. Strube. Mmax: A tool for annotation of multi-modal corpora. In *IJCAI 2001*, pages 45–50, Seattle, 2000.

[18] C. Muller and M. Strube. Multi-level annotation in mmax. In *4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, 2003.

[19] T. A. S. Pardo. *Métodos para Análise Discursiva Automática*. PhD thesis, Universidade de São Paulo, 2005.

[20] T. A. S. Pardo, M. das Graças Volpe Nunes, and L. H. M. Rino. Dizer: An automatic discourse analyzer for brazilian portuguese. In A. L. C. Bazzan and S. Labidi, editors, *Advances in Artificial Inteligence*, volume 3171 of *Lecture Notes in Computer Science*, pages 224–234. Germany: Springer-Verlag, XVII Brazilian Symposium on Artificial Intelligence - SBIA 2004, São Luís, Maranhão, 2004.

[21] D. Rossi, C. Pinheiro, N. Feier, and R. Vieira. Resolução de correferência em textos da língua portuguesa. *Revista Eletrônica de Iniciação Científica*, 1(2), 2001.

[22] S. Salmon-Alt and L. Romary. Data categories for a normalized reference annotation scheme. In *5th International Conference on Discourse Anaphora and Anaphor Resolution*, Furnas, Portugal, September 2004.

[23] E. R. M. Seno and L. H. M. Rino. Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In H. Saggion, editor, *Proceedings of the Workshop on Crossing Barriers in Text Summarization Research*, pages 70–75, 2005.

[24] R. Vieira, C. V. Gasperin, and S. Salmon-alt. *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, volume 1, chapter Coreference and Anaphoric Relations of Demonstrative Noun Phrases in Multilingual Corpus, pages 385–403. John Benjamins, Amsterdam, 1 edition, 2005.

[25] R. Vilela, A. Simoes, E. Bick, and J. J. Almeida. Representacao em xml da floresta sintactica. In J. C. Ramalho, A. Simões, and J. C. Lopes, editors, *XATA2005, XML: Aplicações e Tecnologias Associadas (Vila Verde, Braga, 10 e 11 de Fevereiro de 2005)*, pages 351–361. Universidade do Minho, Fevereiro 2005. http://hdl.handle.net/1822/865.

[26] B. Wing and J. Baldridge. Adaptation of data and models for probabilistic parsing of portuguese. In R. Vieira, P. Quaresma, M. das Gracas Volpe Nunes, N. Mamede, C. Oliveira, and M. C. Dias, editors, *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese PROPOR-06*, Itatiaia, Rio de Janeiro, Brazil, 2006.