UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANOTAÇÃO SEMÂNTICA BASEADA EM ONTOLOGIA: UM ESTUDO DO PORTUGUÊS BRASILEIRO EM DOCUMENTOS HISTÓRICOS DO FINAL DO SÉCULO XIX

JULIANA WOLF PEREIRA

ORIENTADORA: PROFA. DRA. MARILDE TEREZINHA PRADO SANTOS

São Carlos – SP

Julho/2014

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANOTAÇÃO SEMÂNTICA BASEADA EM ONTOLOGIA: UM ESTUDO DO PORTUGUÊS BRASILEIRO EM DOCUMENTOS HISTÓRICOS DO FINAL DO SÉCULO XIX

JULIANA WOLF PEREIRA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software, Banco de Dados e Interação Humano Computador.
Orientadora: Profa. Dra. Marilde Terezinha Prado Santos

São Carlos – SP

Julho/2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

P436as

Pereira, Juliana Wolf.
Anotação semântica baseada em ontologia : um estudo do português brasileiro em documentos históricos do final do século XIX / Juliana Wolf Pereira. -- São Carlos : UFSCar, 2014.
95 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Processamento de textos (Computação). 2. Extração de relações semânticas. 3. Ontologia. 4. Documentos históricos. 5. Mineração de textos. 6. Processamento de linguagem natural (Computação). I. Título.

CDD: 005 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Anotação Semântica baseada em Ontologia: Um estudo do Português Brasileiro em Documentos Históricos do Final do Século XIX"

Juliana Wolf Pereira

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Profa. Dra. Maria de Terezinha Prado Santos
(Orientadora - DC/UFSCar)

Prof. Dr. Ricardo Rodrigues Ciferri
(DC/UFSCar)

Prof. Dr. Fábio Natanael Kepler
(UNIPAMPA)

Prof. Dr. Marcelo Rocha Barros Gonçalves
(UFMS)

São Carlos
Julho/2014

Aos meus pais.

AGRADECIMENTOS

À CAPES, pelo apoio financeiro, e ao Departamento de Computação (DC) da UFSCar, pela estrutura, durante todo o período vigente do mestrado.

A Deus, pela minha vida.

Aos meus pais e irmã por todo amor e apoio.

Ao Marcelo, pela ideia e por toda a inspiração, dedicação e paciência para que esse trabalho se tornasse realidade.

À minha querida orientadora Marilde, pela amizade e carinho, e por toda a sua dedicação e paciência na orientação desde os tempos de graduação. O que me proporcionou motivação e aprendizado imensuráveis.

À Profa. Sandra Abib pela convivência durante a participação no Programa de Mobilidade Acadêmica da ANDIFES - UFMS/UFSCar.

A todos os professores do Grupo de Banco de Dados (GBD), e a todos os professores do Departamento de Computação (DC) da UFSCar.

Aos amigos do LaBDES e do DC, Cristiane, Douglas, Paulo, Fábio, Mirela, Isis, Odair, Thiago, Rafael, Anderson, Mateus, Bruno, Elias, Diego, e muitos outros, pela colaboração e contribuição durante as disciplinas, os almoços no RU, churrascos e cafézinhos.

Aos alunos da UFMS do Câmpus de Coxim, Thasley, Vânia, Ana Paula e Bruno (UNB) pela colaboração na correção e anotação manual dos documentos históricos.

Ao Prof. Dr. Fábio Natanael Kepler e ao Prof. Dr. Ricardo Ciferri, por aceitarem o convite de participação na banca de defesa de mestrado.

“ A degeneração de um povo, de uma nação ou raça, começa pelo desvirtuamento da própria língua.” ... “ Uma raça, cujo espírito não defende o seu solo e o seu idioma, entrega a alma ao estrangeiro, antes de ser por ele absorvida.”

(Rui Barbosa (1849-1923))

RESUMO

Esta dissertação apresenta uma abordagem de anotação semântica automática em documentos históricos do século XIX que discutem a constituição da língua pátria, a Língua Portuguesa no Brasil. O objetivo é gerar um conjunto de documentos semanticamente anotados em acordo com uma ontologia de domínio. Para prover essa ontologia de domínio, foi construída a Ontologia InstrumentoLinguistico que apoiou o processo para a realização da anotação semântica automática. Os resultados obtidos com a anotação foram analisados em comparação com o *Gold Standard* e apresentaram alto grau de coincidência, entre 0.86 e 1.00 para a medida *F1-Score*. Além disso, foi possível localizar novos documentos sobre o domínio discutido em uma amostra das Revistas Brazileiras. Esses resultados comprovam a eficácia da abordagem de anotação semântica automática.

Palavras-chave: Anotação Semântica, Extração de Informação Baseada em Ontologia, Ontologia, Documentos Históricos, Mineração de Texto, Processamento de Língua Natural.

ABSTRACT

This dissertation presents an approach to proceed with semantic annotation in historical documents from the 19th century that discuss the constitution of the mother tongue, the Portuguese Language in Brazil. The objective is to generate a group of semantically annotated documents in agreement with a domain ontology. To provide this domain ontology, the IntrumentoLinguistico Ontology was built, and it supported the process of automatic semantic annotation. The results obtained with the annotation were analyzed in comparison with the Gold Standard and they presented an elevated level of coincidence, between 0.86 and 1.00 for the F1-score measure. Besides that, it was possible to locate new documents about the discussed domain in a sample of the Revistas Brazileiras. These results prove the efficacy of the approach of automatic semantic annotation.

Keywords: Semantic Annotation, Ontology-Based Information Extraction, Ontology, Historical Documents, Text Mining, Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contexto e Motivação

Este trabalho é fruto do desenvolvimento da pesquisa de mestrado na área de Engenharia de Software/Banco de Dados/Interface Homem-Computador no Programa de Pós-Graduação em Ciência da Computação do Departamento de Computação (DC) junto à Universidade Federal de São Carlos (UFSCar). O tópico da pesquisa é multidisciplinar, tendo as áreas Mineração de Texto (MT) na Língua Portuguesa, Processamento de Língua Natural (PLN), Extração de Informação Baseada em Ontologias (OBIE) e Ontologias, como base para discutir o domínio em questão e anotar semanticamente os documentos históricos colhidos em biblioteca digital.

Atualmente as bibliotecas digitais buscam digitalizar coleções raras e históricas para, além da preservação, disponibilizar e divulgar esse material rico e valioso para um público mais abrangente. A norma adotada no Brasil e no exterior para acervos raros não permite que esse material especial faça parte de acervo circulante e que apenas sob condições restritas e controles adequados seja manipulado diretamente por pesquisadores. Nesse sentido, as iniciativas de digitalização são fundamentais para a conservação e preservação desses acervos e, além disso, para tornar livre e irrestrito o acesso a essa informação e documentação pública permitindo, assim, que sejam investigados pelos pesquisadores e também por qualquer outro cidadão interessado.

No Brasil existem algumas iniciativas nesse sentido, como no caso da Hemeroteca Digital Brasileira (HDB)[1]. Parte da Fundação Biblioteca Nacional (FBN)[2], teve sua inauguração em julho de 2012 com mais de 5 milhões de páginas digitalizadas de periódicos brasileiros principalmente dos séculos XIX e XX. A HDB inclui os primeiros jornais brasileiros como, por

[1] http://hemerotecadigital.bn.br/
[2] http://bn.br/

exemplo, o Correio Braziliense (1808), publicações raras como o Jornal das Senhoras (1852) e extintas como o Diário Carioca (1920). Disponibiliza também as edições do período de 1861 a 1979 da Revista Brazileira, esse o objeto de estudo deste trabalho.

O apoio do Ministério da Cultura, Ministério da Ciência, Tecnologia e Inovação e da Financiadora de Estudos e Projetos (FINEP), tornou possível a compra dos equipamentos necessários e a contratação de pessoal para a sua criação e manutenção. A responsabilidade da digitalização do acervo original e produção de microfilmes é do Laboratório de Digitalização FBN. A digitalização dos microfilmes foi executada pela DocPro, essa empresa também desenvolveu a tecnologia DocReader para realizar busca textual dentro dos acervos.

Outro exemplo no cenário nacional é o da Brasiliana USP[3]. A Biblioteca Brasiliana Guita e José Mindlin possui o acervo doado em 2006 pelo casal Mindlin, configurando-se como a maior coleção de livros e documentos que falam sobre o Brasil, tanto escritos por brasileiros quanto por estrangeiros. Em 2009, foi estabelecido um laboratório de digitalização e instalado um digitalizador robotizado na casa dos colecionadores para o início da digitalização dos livros. Além disso, aplicativos foram projetados e implementados baseados em software livre para sustentar uma biblioteca digital.

Com forte apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), do Ministério da Cultura, da Petrobras e do Banco Nacional de Desenvolvimento Econômico e Social (BNDES), o Laboratório Brasiliana USP tornou-se referência nacional no campo da digitalização de acervos. No ano de 2013, passou a ocupar parte do edifício da Biblioteca Mindlin com novos sistemas robotizados de digitalização e servidores que permitiram atender à demanda computacional exigida para a criação de repositórios e metabuscadores para o acervo.

Após o processo de digitalização desses documentos, é necessário que os arquivos digitais passem por um processo de reconhecimento automático de caracteres *(Optical Character Recognition - OCR)*. A finalidade é de transformar os arquivos de imagem em sequências de caracteres interpretáveis por programas de computador e permitir a preparação desses documentos para extração de informações. Essa transformação proporciona ir além da visualização da imagem escaneada do documento, ela permite o acesso computacional ao conteúdo do documento possibilitando, assim, que sejam aplicadas várias tecnologias de processamento de língua natural.

No entanto, podemos observar nos trabalhos de Paixão De Sousa, Kepler e Faria (2010) e Aluísio (2007), uma discussão sobre os desafios no tratamento de documentos históricos. Devem ser observados os erros originários do OCR porque esse tipo de documento apresenta carac-

[3]http://www.bbm.usp.br/

terísticas muito peculiares, como junção de palavras, abreviaturas, variação de grafia, alterações grafemáticas, etc., como perceberam os autores analisando documentos escritos entre os anos 1500 e 1808 no córpus do Dicionário Histórico do Português do Brasil (CÂNDIDO JR; ALUÍSIO, 2009) (VALE et al., 2008) (GIUSTI et al., 2007) e no córpus Tycho Brahe (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010) (PAIXÃO DE SOUSA; TRIPPEL, 2006).

A variação de grafia permanece nos textos escritos durante o século XIX e são encontradas nos textos ocorrências de uma grande variação gráfica (acentuação, consoantes dobradas) como é o caso da Revista Brazileira. Os trabalhos que lidam com as peculiaridades de textos históricos são necessários, principalmente, para garantir as características originais dos documentos e permitir uma adequada análise linguística e filológica. Essa discussão é de fundamental importância no contexto dos documentos históricos, porém, o foco desta dissertação de mestrado está em lidar com os documentos já tratados.

Vários trabalhos no Brasil, como a HDB e a Brasiliana, disponibilizam o acervo digitalizado a todos livremente, com acesso irrestrito e com a possibilidade de baixar os documentos para o computador. No entanto, os sistemas de busca envolvidos restringem-se à recuperação por metadados e palavra-chave.

Com o intuito de criar elementos semânticos que podem ajudar a melhorar futuramente esses sistemas de busca, nesta dissertação é apresentada uma abordagem para anotação semântica baseada em ontologia num estudo do Português Brasileiro em documentos históricos do final do século XIX.

Nesse período, como afirma Gonçalves (2012), além das variações gráficas, os conceitos de língua e dialeto apresentam certa instabilidade, revelando posições distintas de brasileiros e portugueses em relação à constituição da língua portuguesa no Brasil. Necessitamos levar em conta as relações entre os conceitos de língua portuguesa, língua nacional, dialeto, e outros, para possibilitar melhores recursos para a busca de textos nesse período. Não é sempre, por exemplo, que dialeto está subordinado à língua, expressando uma relação de inferioridade. Especialmente no caso do Brasil, nem mesmo se pode falar categoricamente que houve o alçamento de uma variedade regional à posição de língua oficial.

Na Revista Brazileira, em sua fase que compreende o período de 1879 a 1900, é predominante a discussão da diferença da Língua Portuguesa no Brasil em relação a Portugal. Esse conjunto de textos reunidos em torno de temáticas tão distintas (lexicologia, sintaxe, fonologia, etc.), mas de alguma maneira convergentes no sentido de separar as duas línguas em questão, pode exemplificar o intenso trabalho de instrumentação da língua no Brasil durante este período, reflexo assim da gramatização (AUROUX, 1992) pela qual passava a língua portuguesa. São

comuns, por exemplo, a utilização dos paradigmas verbais, das traduções interlineares, dos conjuntos de regras e exemplos (GONÇALVES, 2012).

A mais recente compilação de textos da Revista Brazileira sobre a temática da Língua Portuguesa foi publicada em fac-símile pela Academia Brasileira de Letras em 2005 e teve seu prefácio assinado por Evanildo Bechara (BECHARA, 2005).

1.2 Hipóteses e Objetivos

O objetivo desse trabalho é prover anotação semântica automática em documentos históricos do século XIX que discutem a constituição da língua pátria, a Língua Portuguesa no Brasil. Para dar apoio a esse processo foi necessária a construção de uma ontologia de domínio. Além disso, foi desenvolvido um *Gold Standard* para avaliação da abordagem de anotação semântica automática. Neste contexto, essa dissertação apresenta uma abordagem de anotação semântica automática apoiada na Ontologia InstrumentoLinguistico, com objetivo de anotar os documentos da Revista Brazileira.

Sob essa perpectiva, tem-se por hipóteses:

- É possível realizar uma anotação semântica nos documentos históricos para identificar quais ocorrências abordam o tema investigado.
- Uma ontologia de domínio colabora para uma efetiva anotação semântica produzida por um processo de extração de informação baseada em ontologia.
- É possível definir uma ontologia de domínio que expressa o conhecimento da discussão da constituição da Língua Portuguesa no Brasil.
- É possível anotar semanticamente os documentos sem realizar a análise morfossintática no córpus.

1.3 Organização do Trabalho

Esta dissertação está organizada da seguinte maneira:

- Capítulo 2: são apresentados os conceitos necessários para o entendimento da abordagem desenvolvida.

- Capítulo 3: são apresentados os trabalhos correlatos existentes na literatura.
- Capítulo 4: é apresentada a metodologia para a construção e a documentação da Ontologia InstrumentoLinguistico.
- Capítulo 5: é apresentada a abordagem de anotação semântica automática baseada em ontologia. A arquitetura geral da abordagem, o Pré-Processamento, a Construção dos Artefatos e a Extração de Informação baseada em ontologia.
- Capítulo 6: é apresentada a comparação com o *Gold Standard*, a discussão dos resultados e a avaliação da abordagem.
- Capítulo 7: é apresentada a conclusão desta dissertação e os trabalhos futuros.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2.1 Mineração de Texto

A mineração de texto busca extrair informações em uma coleção de dados não estruturados ou semiestruturados de maneira análoga à mineração de dados (FELDMAN; SANGER, 2007). O processo de descoberta de conhecimento em dados textuais, também definido como *Knowledge Discovery in Text* (KDT), utiliza um conjunto de ferramentas de análise para identificar e explorar padrões e informações úteis, interessantes e não triviais (FELDMAN; DAGAN; HIRSH, 1998).

A utilização de recursos que apoiam a validação, a normalização ou cruzamento de características dos documentos como, por exemplo, dicionários, ontologias ou bases de conhecimento, podem tornar computacionalmente mais eficientes a identificação das características e facilitar a descoberta de padrões, para auxiliar na geração de menores conjuntos de características, porém mais ricos semanticamente. O conhecimento de domínio fornecido pelas ontologias e bases de conhecimento, permitem representações sofisticadas de hierarquias conceituais, sendo assim, representam características mais semânticas do que outros tipos de recursos (FELDMAN; SANGER, 2007).

2.1.1 Processamento de Língua Natural

As tarefas de Processamento de Línguas Naturais (PLN) podem ser aplicadas para pré-processar os documentos, a fim de prepará-los para futuros processos. Entre outras tarefas, podem ser utilizadas a Tokenização, a Sentenciação e o *PosTagger*.

Tokenização

A tokenização é o processo que segmenta o documento em partes denominadas de *tokens*. Os *tokens* podem ser palavras, números, pontuação e outros, desde que seja uma unidade semântica útil para processamento. Uma palavra é uma sequência contínua de caracteres alfanuméricos, separada por espaços em branco ou por caracteres de pontuação. Da mesma forma, os números são definidos como uma sequência consecutiva de dígitos. Tanto a palavra, quanto os números, são marcados com uma marcação *Token* que os identifica do primeiro ao último caractere. Os caracteres de pontuação são reconhecidos individualmente e marcados com a marcação *Token*. Os espaços em branco são marcados com a marcação *SpaceToken*.

Sentenciação

A Sentenciação é o processo que segmenta o documento em sentenças. A sentença pode ser identificada quando é detectado um caractere de pontuação que marca o fim de uma sentença. Cada sentença é anotada com a marcação *Sentence*. Cada ponto final é marcado com uma marcação *Split*.

PosTagger

A anotação morfossintática *(PosTagger)* marca os *tokens* com o seu tipo de palavra correspondente com base no próprio *token* e no contexto do *token*. Um *token* pode ter várias marcações *pos* dependendo do *token* e do contexto. Para limitar as marcações possíveis para um *token*, um dicionário pode ser utilizado, porém, isso aumenta o tempo de execução da marcação e desempenho do PosTagger.

2.1.2 Extração de Informação Baseada em Ontologia

A Extração de Informação Baseada em Ontologia (OBIE) é um tema recente de pesquisa na subárea da Extração de Informação (EI) que é uma subárea de Processamento de Língua Natural. Esse tipo de sistema processa qualquer tipo de documento não estruturado (arquivos de texto) ou semiestruturado (HTML, XML) escrito em linguagem natural, com técnicas de EI guiadas por ontologias para extrair conceitos, propriedades e relações expressas em uma ontologia (SAGGION et al., 2007).

Isto é, utiliza os métodos existentes de EI apoiados por uma ontologia para localizar nos documentos o conhecimento representado pela ontologia. A informação pode ser exportada como uma ontologia com instâncias, ou as anotações nos textos contêm *links* para a ontologia.

De acordo com Wimalasuriya e Dou (2010b), as ontologias desempenham um papel fundamental no processo de extração de informação. Elas formalizam e estruturam o domínio para agilizar a extração de conhecimento em documentos que são semanticamente anotados e acessados utilizando o vocabulário fornecido pela ontologia (IRIA et al., 2004).

Os principais métodos de EI utilizados são regras linguísticas representadas por expressões regulares, dicionário, técnicas de classificação, construção de árvores de *parsers*, análise de marcações HTML/XML e buscas baseadas na web (WIMALASURIYA; DOU, 2010b).

A principal diferença de um sistema OBIE e a Extração de Informação, é a utilização de uma ontologia formal como entrada do sistema, ao invés de apenas um léxico ou dicionário. Além disso, como saída a entidade extraída está vinculada à sua descrição semântica na ontologia (MAYNARD et al., 2005). No entanto, essa ontologia também pode ser construída com os resultados obtidos a partir das técnicas de extração de informação (WIMALASURIYA; DOU, 2010a).

Se uma ontologia de domínio pode ser utilizada com sucesso por um sistema OBIE para extrair os conteúdos semânticos de um conjunto de documentos relacionados a esse domínio, pode-se deduzir que a ontologia é uma boa representação do domínio (WIMALASURIYA; DOU, 2010b). Além de contribuir potencialmente para o desenvolvimento da Web Semântica, pois está relacionada com a representação do conhecimento.

2.1.3 Anotação Semântica

O processo de vincular modelos semânticos e linguagem natural é conhecido como anotação semântica. Esse processo cria inter-relações entre ontologias e documentos não estruturados ou semiestruturados (LI; BONTCHEVA, 2007). A anotação semântica é a atribuição de *links* para a descrição semântica de cada entidade localizada nos documentos (KIRYAKOV et al., 2004).

Segundo Uren et al. (2006) as ferramentas que permitem a produção de anotações semânticas podem ser manuais ou automáticas. A anotação manual é feita por ferramentas que fornecem um editor de texto e suporte para ontologias, podendo possuir recursos que apoiam a anotação semântica, como é o caso da GATE. A anotação semiautomática pode ser feita por ferramentas que proporcionam sugestões para as anotações, porém, que ainda necessitam da intervenção de especialistas. As anotações automáticas podem ser feitas por ferramentas que realizam as anotações automaticamente. Geralmente esses processos são apoiados em técnicas de extração de informação baseadas em ontologias.

Um OBIE utiliza de técnicas Extração de Informação adaptadas para realizar a tarefa de

anotação semântica (LI; BONTCHEVA, 2007).

2.2 Ontologias

Ontologias tem sido empregadas para a representação do conhecimento, principalmente com o desenvolvimento da Web Semântica (BERNERS-LEE et al., 2001), que investiga como tratar a semântica do conteúdo e de serviços disponíveis na Web.

Por representarem informação semântica e permitirem inferência de conhecimento, ontologias tem sido utilizadas em diversas aplicações para facilitar a comunicação tanto entre humanos quanto entre os sistemas computacionais (YAGUINUMA; SANTOS; BIAJIZ, 2007).

Nos sistemas OBIE as ontologias proporcionam a estrutura conceitual de um domínio específico, apoiando o processo de anotação semântica dos documentos. Os sistemas de Recuperação de Informação podem ser especialmente favorecidos, pois essas anotações podem atuar como índice e facilitar a busca semântica e a recuperação de conhecimento em documentos (KIRYAKOV et al., 2004) (BREWSTER, 2002).

Pela definição de Gruber (1993), Borst (1997) e Studer, Benjamins e Fensel (1998): "Uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada". Formal pois deve permitir a interpretação por máquinas. Explícita, pois conceitos, propriedades, relacionamentos e axiomas devem ser explicitamente definidos. A conceitualização é o modelo abstrato do mundo real. E é compartilhada, pois esse conhecimento é consensual.

Para a Ciência da Computação, uma ontologia define um conjunto de primitivas representacionais com as quais é possível modelar um domínio de conhecimento ou discurso (GRUBER, 2009). De modo geral, tais primitivas são hierarquias de conceitos (taxonomias), atributos, relacionamentos, indivíduos e restrições (axiomas), que descrevem o conhecimento de um determinado domínio de modo consensual e compartilhado (GUARINO, 1998).

De acordo com Noy e Mcguinness (2001), uma ontologia é uma descrição formal de conceitos ou classes em um domínio de discurso. As classes podem estar organizadas na forma de uma taxonomia, permitindo subclasses e superclasses. As propriedades descrevem as características e os atributos de cada conceito, podem ser divididas em propriedades de objetos *(Object Property)* e propriedades de tipos dados *(DataType Property)*. A primeira determina relações binárias entre indivíduos ou classes. A segunda relaciona um indivíduo a um dado literal. As restrições sobre os conceitos e propriedades expressam significado de forma que máquinas sejam capazes de interpretar por meio de raciocínio automático. Por fim, os in-

divíduos ou instâncias de classes, junto à ontologia, constituem a base de conhecimento.

2.2.1 Representação de Ontologias

RDF *(Resource Description FrameWork)* representa um modelo de armazenamento de dados onde as informações são expressas em forma de triplas (KLYNE; CARROLL, 2006) (PAN, 2009). A tripla é formada por sujeito, predicado e objeto, na Figura 2.1 está representada a estrutura de uma tripla. Cada item da tripla é representado por uma URI *(Uniform Resource Identifier)* que identifica unicamente cada recurso de uma ontologia. É uma camada que fornece uma estrutura básica para o compartilhamento de informações, no entanto, a semântica é limitada.

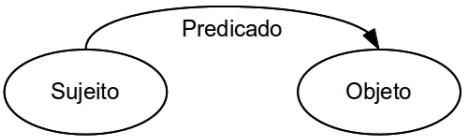


Figura 2.1: Estrutura de uma tripla RDF.

Com o RDFS *(RDF Schema)* a semântica é um pouco menos limitada, pois permite a definição de classes, recursos e propriedades (PAN, 2009). Permite, ainda, definir uma hierarquia entre as propriedades e também estipular quais classes podem relacionar-se com elas, por meio do uso de domínios e alcance *(domain, range)*.

A partir da OWL *(Web Ontology Language)* (MCGUINNESS; HARMELEN et al., 2004), recomendação da W3C *(World Wide Web Consortium)*, é possível definir igualdade e desigualdade entre classes, características para as propriedades, restrições de propriedades e restrições de cardinalidade, garantindo assim uma semântica bem mais expressiva. As características de propriedades são:

- Inversa: Uma propriedade pode ser indicada como a inversa de outra propriedade. Se a propriedade P1 é indicada para ser a inversa da propriedade P2, então, uma instância x está relacionada com uma instância y pela propriedade P2, consequentemente, y está relacionada com x pela propriedade P1.
- Transitiva: Propriedades podem ser transitivas. Se uma propriedade é transitiva, então o par (x, y) é uma instância da propriedade transitiva P, e o par (y, z) é uma instância de P, consequentemente, o par (x, z) também é uma instância de P.

- Simétrica: Propriedades podem ser simétricas. Se uma propriedade é simétrica, então o par (x, y) é uma instância da propriedade simétrica P, e o par (y, x) também é uma instância de P.
- Funcional: Se uma propriedade é funcional pode-se afirmar que tem um valor único, uma instância x só pode estar conectada a no máximo uma instância y.
- Funcional inversa: Se uma propriedade é funcional inversa então o inverso da propriedade é funcional. Assim, o inverso da propriedade tem no máximo um valor para cada instância. Então, um instância y só pode estar relacionada com uma instância x.

2.2.2 SPARQL

O SPARQL (PRUD'HOMMEAUX; SEABORNE et al., 2008) é uma linguagem de consulta e um protocolo para acesso a dados armazenados em RDF. Foi desenvolvido para consultar RDF, por isso a consulta é descrita em padrões de triplas similares as do RDF. Permite realizar consultas na OWL pois esta linguagem é baseada na linguagem RDF. Em constante desenvolvimento, a última especificação da W3C é a versão 1.1 lançada em 2013, permite além das consultas, inserir, atualizar e remover triplas. Essas consultas podem ser executadas no Protégé no painel *SPARQL Query*. As consultas foram utilizadas para auxiliar na construção dos artefatos para a abordagem de anotação semântica automática.

2.3 Plataformas de Desenvolvimento

2.3.1 Protégé

A plataforma Protégé[1] é um ambiente de desenvolvimento que provê suporte a um conjunto de ferramentas *(plugins)* para a construção de modelos de domínio e aplicações baseadas em conhecimento com ontologias. Atualmente encontra-se na versão *Protégé Desktop 5.0 beta*, é implementada em Java e *open-source*.

Um exemplo de *plugin* é o editor Protégé-OWL que permite a criação e edição de ontologias atráves de uma interface gráfica, facilitando o processo de construção e a manipulação de classes, instâncias, propriedades e axiomas em um projeto de ontologia. Gera o código em linguagem OWL ou RDF/XML da ontologia desenvolvida, além de exportar para formatos como *Turtle*, *N-Triples*, *N3* e *Clips*.

[1] http://protege.stanford.edu/

Outros exemplos são os *plugins* para visualização da ontologia, como o OntoViz[2] e o OWLViz[3], e para a realização de consultas como o *SPARQL Query*[4]. Além desses, existem muitos outros para diversas funcionalidades, e também é possível desenvolver *plugins* personalizados que atendam algum interesse específico.

A versão *Protégé 3.5* foi utilizada com os *plugins* Protégé-OWL, OntoViz e SPARQL Query, para auxiliar na construção da Ontologia InstrumentoLinguistico.

2.3.2 Gate

O GATE *(General Architecture for Text Engineering)*[5] é uma plataforma *open-source* de desenvolvimento integrado para processamento de língua natural, é implementada em Java e sua idealizadora é a Universidade de Sheffield (CUNNINGHAM; WILKS; GAIZAUSKAS, 1996) (CUNNINGHAM et al., 2014). Permite ao desenvolvedor criar seus próprios recursos ou estender os existentes, e construir aplicações para o processamento de língua natural. Dispõe de um sistema de extração de informação, *A Nearly-New Information Extraction System (ANNIE)*, e possui um conjunto abrangente de *plugins (CREOLE)* proporcionando a integração com ferramentas como Kea, LingPipe, WordNet, OpenNLP, que atendem a diferentes funcionalidades. Provê o suporte a documentos de diversos formatos, por exemplo, XML, PDF, MS Word, OpenOffice Writer, TXT, email, permite realizar anotação manual, extração de informação, anotação semântica, entre outras tarefas.

Os recursos ou componentes podem ser de três tipos:

- Recursos de Linguagem - como córpus ou ontologia.
- Recursos de Processamento - algoritmos como tokenizador, sentenciador, regras *JAPE*.
- Recursos de Visualização - visualização e edição dos recursos na interface gráfica.

Nesta dissertação são utilizados como recursos de linguagem um córpus e uma ontologia de domínio. Como recursos de processamento um tokenizador *OpenNLP*, um dicionário e uma lista de mapeamento *OntoGazzeter* e um conjunto de regras *JAPE*.

[2] http://protegewiki.stanford.edu/wiki/OntoViz
[3] http://protegewiki.stanford.edu/wiki/OWLViz
[4] http://protegewiki.stanford.edu/wiki/SPARQL_Query
[5] http://gate.ac.uk/

OpenNLP

O OpenNLP fornece um conjunto de ferramentas baseadas em aprendizado de máquina para processar textos em linguagem natural. Essas ferramentas são implementadas em Java e utilizam as bibliotecas OpenNLP *Tools* e OpenNLP *Maxent*. Servem para realizar tokenização, sentenciação, anotação PosTagger, entre outras tarefas[6]. Os modelos estão disponibilizados para diversas línguas, incluindo os modelos para a Língua Portuguesa utilizados nesta abordagem[7].

JAPE

O JAPE *(Java Annotation Patterns Engine)* fornece uma gramática para a criação de regras que possibilitam realizar anotações sobre padrões localizados nos documentos. Essas regras são implementadas como um conjunto de transdutores de estado finito e cada transdutor geralmente contém uma regra que determina um tipo diferente (CUNNINGHAM; MAYNARD; TABLAN, 2000)(MAYNARD et al., 2005).

A gramática consiste em um conjunto de fases que são formadas por uma ou mais regras. O arquivo denominado *main.jape* determina a sequência em que essas fases são executadas, criando uma cascata de transdutores de estados finitos.

Cada regra possui um lado esquerdo *(LHS)*, antes da seta, e um lado direito *(RHS)*, depois da seta. O lado esquerdo consiste na descrição do padrão a ser buscado para a anotação e pode conter operadores de expressão regular, por exemplo, (*) para determinar zero ou mais ocorrências, (+) para determinar uma ou mais ocorrências e (|) para determinar o "ou". O lado direito consiste na manipulação ou ação sobre as anotações atráves de código em Java. Além disso, nas especificações das regras podem ser determinadas prioridades e controles sobre as anotações (CUNNINGHAM; MAYNARD; TABLAN, 2000).

ANNIE

O ANNIE *(A Nearly-New Information Extraction System)* é o sistema de extração de informação distribuido junto ao GATE, sua implementação é baseada em algoritmos de estado finito e na linguagem JAPE. Disponibiliza inúmeros componentes como, por exemplo, *Tokenizer*, *Sentence Splitter* e *OrthoMatcher* que, por padrão, são para a língua inglesa e podem ser combinados para extrair entidades. Além disso, pode ser adaptado com outros recursos para apoiar diferentes tipos de processos de extração de informação para outras línguas (CUNNINGHAM et al., 2014).

[6] http://opennlp.apache.org/

[7] http://opennlp.sourceforge.net/models-1.5/

Annotation Diff Tool

Com essa ferramenta é possível comparar dois conjuntos de anotação do mesmo tipo em um ou dois documentos. Todas as anotações do conjunto chave *(Key)* são comparadas com o conjunto resposta *(Response)*. Para cada tipo de anotação são calculados precisão, revocação e F-Measure. O *Corpus Quality Assurance* estende as funcionalidades de *Annotation Diff* para analisar todo o córpus através de uma interface amigável (CUNNINGHAM et al., 2014).

Ontologias

No GATE as ontologias são classificadas como recursos linguísticos e para criá-los são necessários dois *plugins*. Esses *plugins* fornecem diversas ferramentas para edição de ontologias e anotação de documentos utilizando a ontologia. O *plugin "Ontology"* fornece uma *API* para a manipulação de ontologias. O *plugin "Ontology Tools"* fornece os recursos OntoGazzeter, um editor para ontologias e o OAT que permite realizar a anotação semântica manual nos documentos (CUNNINGHAM et al., 2014).

OntoGazzeter

O recurso de processamento OntoGazzeter permite associar cada entidade presente nas listas que formam o dicionário com uma classe na ontologia de domínio, através de uma lista de mapeamento fornecida pelo desenvolvedor. As listas que formam o dicionário são arquivos de texto simples com uma entrada por linha. A lista de mapeamento associa cada lista com a classe correspondente na ontologia de domínio.

2.4 Medidas de Desempenho

As métricas Precisão, Revocação e F-Measure, instrínsecas no campo de Recuperação de informação (FRAKES; BAEZA-YATES, 1992), podem também ser utilizadas na Extração de Informação para medição de desempenho (COWIE; LEHNERT, 1996). E na Extração de Informação Baseada em Ontologia (MAYNARD; PETERS; LI, 2006).

Nesta dissertação essas medidas são utilizadas em dois momentos. No primeiro momento, na metodologia de construção da ontologia, para comparar as anotações realizadas manualmente por um grupo de anotadores em relação ao *Gold Standard* inicial feito pelo especialista de domínio. Em segundo momento, para avaliar as anotações provenientes da abordagem de anotação semântica automática em relação ao *Gold Standard* final, conforme apresentada no Capítulo 6 onde ocorrem as discussões dos resultados deste trabalho.

De acordo com Wimalasuriya e Dou (2010b), as medidas de desempenho podem ser representadas pelas seguintes fórmulas:

$$Precisão = \frac{Anotações\ Corretas}{Total\ de\ Anotações} \tag{1}$$

$$Revocação = \frac{Anotações\ Corretas}{Total\ de\ Anotações\ no\ Gold\ Standard} \tag{2}$$

Precisão (Eq. 1) mede o número de ocorrências corretamente anotadas como uma porcentagem do número de ocorrências anotadas. Em outras palavras, mede quantas das ocorrências que o anotador ou a aplicação identificaram são realmente corretas, independentemente de não terem conseguido recuperar todos as ocorrências corretas. Quanto maior a precisão, melhor é a anotação e assegura que o que foi anotado é correto.

Revocação (Eq. 2) mede o número de ocorrências corretamente anotadas como uma porcentagem do número total de ocorrências no *Gold Standard*. Em outras palavras, mede quantas das ocorrências que deveriam ter sido anotadas foram realmente anotadas, independentemente de quantas anotações incorretas foram feitas. Quanto maior a taxa de revocação, melhor o desempenho do anotador ou da aplicação nas anotações de ocorrências corretas.

A medida F-Measure (Eq. 3) é a média ponderada da conjugação da precisão e revocação.

$$F-Measure = \frac{(\beta^2+1)^*Precisão^*Revocação}{(\beta^{2*}Precisão)+Revocação} \tag{3}$$

Onde β determina o peso da precisão *versus* a revocação. Geralmente $\beta = 1$, o que resulta no mesmo peso para a precisão e revocação (Eq.4). Considerando isso, a medida também pode ser denominada como F1-Score.

$$F1-Score = \frac{2^*Precisão^*Revocação}{Precisão+Revocação} \tag{4}$$

Cada uma das medidas descritas acima podem ser calculadas com dois critérios diferentes, *Strict* e *Lenient*, a fim de considerar *match* exato e *match* aproximado. *Strict* considera todas as respostas parcialmente corretas como incorreto. *Lenient* considera todas as respostas parcialmente corretas como correto.

Capítulo 3

TRABALHOS CORRELATOS

3.1 Trabalhos desenvolvidos para o tratamento de Documentos Históricos do Português do Brasil

O Dicionário Histórico do Português do Brasil (DHPB) foi idealizado pela Profa. Maria Tereza Camargo Biderman. Registra as mudanças semânticas que as unidades lexicais sofrem ao longo do tempo e é relativa aos séculos XVI, XVII e XVIII (MURAKAWA, 2009).

O trabalho de Cândido Jr e Aluísio (2008) é um ambiente computacional para processamento de córpus, criação de glossários e redação de verbetes para o DHPB, sendo possível aplicá-lo a outros projetos de criação de dicionários históricos.

O Tycho Brahe "Corpus Histórico do Português Tycho Brahe" é um córpus eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1845 (GALVES; FARIA, 2010). No trabalho de Menegatti (2002) são apresentadas regras linguísticas para o tratamento computacional da variação de grafia e abreviaturas do córpus Tycho Brahe. Diferentes abordagens são apresentadas por Hirohashi (2004) para detecção automática de variações de grafia, como por agrupamento por distância de edição, análise fonética e regras de normalização aprendidas automaticamente. O E-Dictor é uma ferramenta auxiliar para edição eletrônica em XML de documentos históricos com vistas a análise linguísta automática, criado a partir de demandas na construção do Corpus Anotado do Português Tycho Brahe (CTB) (SOUSA; KEPLER; FARIA, 2009).

Esses trabalhos proporcionaram um grande avanço nos métodos e ferramentas para análise de documentos históricos. O que difere nossa abordagem das citadas anteriormente é que nosso objetivo é anotar semanticamente documentos históricos.

3.2 Trabalhos desenvolvidos para o tratamento de Documentos Históricos em outras línguas

No trabalho de Ernst-Gerlach e Fuhr (2007), é apresentada uma abordagem para recuperação em documentos históricos escritos na Língua Alemã, em que a ortografia não é padronizada. Neste trabalho são analisadas as variações morfológicas e a construção de regras para variantes ortográficas. O termo de pesquisa é tratado como um lema e é utilizado um dicionário contemporâneo de alemão para encontrar todas as inflexões e derivações desse lema. Em seguida são aplicadas regras de transformação para gerar as variantes ortográficas históricas, essas regras derivam de dados de treinamento. Assim, quando o usuário digita um termo de pesquisa (lema) o dicionário produz todas as formas da palavra. Logo após, são geradas as variantes de ortografia destas formas. Os resultados experimentais mostram importante melhora na qualidade de recuperação em coleções históricas.

Outra abordagem, discutida por Kempken, Luther e Pilz (2006), mostra como recuperar informação em uma coleção de documentos escritos no século XVII em Língua Holandesa. São avaliadas as medidas de distância fonética, correção dos erros de digitação e similaridade entre strings. Os autores comprovam que as consultas modernas não são eficazes para a recuperação de documentos históricos, no entanto, ferramentas específicas para a linguagem histórica aumentam a eficácia da recuperação. As melhorias são significativas e estão muito além ao uso de algoritmos modernos decorrentes.

No trabalho de Rayson, Archer e Smith (2005), é apresentada uma abordagem para lidar com ortografia histórica da Língua Inglesa. Os autores desenvolveram um detector de variâncias (VARD) para textos escritos em inglês entre os século XVI e XIX. É preciso destacar também que elevaram a precisão por meio da normalização das variantes históricas para o contemporâneo.

3.3 Trabalhos desenvolvidos para construção do Gold Standard

Trabalhos que envolvem a construção, a avaliação e o refinamento de um *Gold Standard* são propostos para o domínio biomédico. Na proposta de Roberts et al. (2009) um córpus semanticamente anotado é construído para desenvolver e avaliar sistemas de extração de informação em registros de pacientes.

Outra abordagem (VELUPILLAI et al., 2009), discute a criação de dois *Gold Standards*, um

criado automaticamente, e um criado a partir das discussões entre os anotadores. Os dados utilizados são um subconjunto do *Stockholm EPR Corpus*. As entidades e relacionamentos que formam o modelo de anotação tiveram como base uma ontologia que modela todos os aspectos do paciente e do tratamento nos documentos clínicos. O corpus resultante é um rico recurso semanticamente anotado para processamento de texto clínico.

Esses trabalhos influenciaram nos procedimentos adotados em nossa abordagem para a criação do *Gold Standard*.

3.4 Trabalhos de OBIE aplicados em outros domínios

São apresentados na Tabela 3.1, conforme classificação de Wimalasuriya e Dou (2010b), alguns trabalhos de OBIE aplicados em outros domínios. Nela são comparados, nos trabalhos citados e na nossa abordagem, os métodos de extração de informação, a forma como foi construída a ontologia, os componentes extraídos da ontologia, bem como os tipos de documentos utilizados.

Tabela 3.1: Comparação de sistemas de OBIE, adaptada de Wimalasuriya e Dou (2010b)

Sistema OBIE	**Métodos de Extração de Informação**	**Componentes extraídos da Ontologia**	**Tipos de Documentos**
Saggion et al. (2007)	Regras Linguísticas, listas *Gazetteer*	Instâncias, valores de propriedade	Documentos de um domínio
Li e Bontcheva (2007)	Classificação	Classes, instâncias	Documentos de um domínio
Vargas-Vera et al. (2001)	Regras Linguísticas	Instâncias, valores de propriedade	Páginas web de um site particular
Popov et al. (2004)	Regras Linguísticas, listas *Gazetteer*	Instâncias, valores de propriedade	Documentos de um domínio
Nossa Abordagem	Regras Semânticas, Dicionário (listas *Gazetteer*)	Classes, instâncias, valores de propriedade, taxonomia, outros relacionamentos	Documentos de um domínio

Do ponto de vista do método de extração de informação apenas Li e Bontcheva (2007) não utilizam regras. Vale lembrar que na nossa abordagem as regras são semânticas, ao passo que nas outras as regras são linguísticas, e que denominamos as listas *Gazetteer* como dicionário.

Como na nossa abordagem, todas as demais propostas utilizaram documentos de um domínio

específico, com exceção de Vargas-Vera et al. (2001) que utilizam páginas web de um site particular.

O diferencial da nossa abordagem está nos componentes extraídos da ontologia que vão além de classes e instâncias, permitindo extrair valores de propriedades, taxonomia e outros relacionamentos.

Capítulo 4

ONTOLOGIA INSTRUMENTOLINGUISTICO

4.1 Metodologia para Construção da Ontologia InstrumentoLinguistico

O processo de construção da Ontologia InstrumentoLinguistico é iterativo e a cada ciclo a ontologia é refinada e aperfeiçoada. Para apoiar a construção da Ontologia InstrumentoLinguistico é utilizada a metodologia desenvolvida por Uschold e King (1995). Conforme o estabelecido pelos autores, as seguintes fases formam a linha mestra da metodologia:

- Identificação do propósito.
- Construção da Ontologia.
  - Análise Ontológica.
  - Codificação da Ontologia.
  - Integração com ontologias existentes.
- Avaliação.
- Documentação.

É importante identificar o propósito da construção da ontologia e para que ela será utilizada. Nesta dissertação, a Ontologia InstrumentoLinguistico estrutura uma base de conhecimento sobre o domínio da Constituição da Língua Portuguesa no Brasil e tem como intuito apoiar o processo de anotação semântica em documentos históricos.

Para a construção da ontologia, os autores sugerem que deve ser feita a análise ontológica do domínio para identificar os principais termos, conceitos e relações no domínio de interesse, isto é, qual o escopo investigado. Como sugestão, os autores incentivam realizar um *Brainstorming*

na intenção de produzir um conjunto de termos e frases que representam o domínio. Pensando nisso, para apoiar a descoberta de conhecimento sobre o domínio e reproduzir o *Brainstorming* sugerido, realizou-se primeiramente a construção do córpus com os documentos indicados pelo Especialista de Domínio e, posteriormente, um processo de anotação manual efetuado por anotadores do domínio, ambos estão descritos detalhadamente nas próximas seções.

A representação formal da Ontologia InstrumentoLinguistico, para representar explicitamente a conceitualização identificada, foi realizada utilizando a linguagem *OWL*. O código foi implementado pelo Engenheiro de Ontologia com o auxílio da ferramenta *Protégé*.

Como não foram localizadas ontologias que representam o domínio discutido, a etapa de integração com ontologias existentes não é realizada. Uma avaliação inicial foi realizada pelo especialista de domínio, porém, uma etapa de avaliação mais ampla realizada por outros especialistas é proposta como trabalho futuro desta dissertação de mestrado.

Alguns extratos da documentação da Ontologia InstrumentoLinguistico são apresentados na Seção 4.2.

4.1.1 Construção do córpus

O córpus, denominado **Córpus Revista Brazileira**, é formado por um conjunto de cinco documentos da Revista Brazileira durante a fase Midosi e discute a problemática da diferença da Língua Portuguesa no Brasil em relação a Portugal. Os documentos, cujos dados estão contidos na Tabela 4.1, são investigados e discutidos a partir de uma leitura crítica dos documentos em (GONÇALVES, 2012), pois apresentam discussões em torno do domínio da Constituição da Língua Portuguesa no Brasil. Ainda que Gonçalves (2012) perceba que existam relacionamentos entre termos e conceitos dentro desses documentos, não há na pesquisa uma modelagem computacional desse domínio.

Para tornar possível uma modelagem do conhecimento desse domínio, é necessária a realização de um pré-processamento nos documentos da amostra que consiste na conversão do formato PDF para um formato ODT ou DOC. Esta conversão permite a correção dos erros de OCR, a fim de garantir que esses documentos não apresentem erros desse tipo que podem interferir na análise para aquisição do conhecimento.

Pré-processamento para preparação dos Documentos do córpus

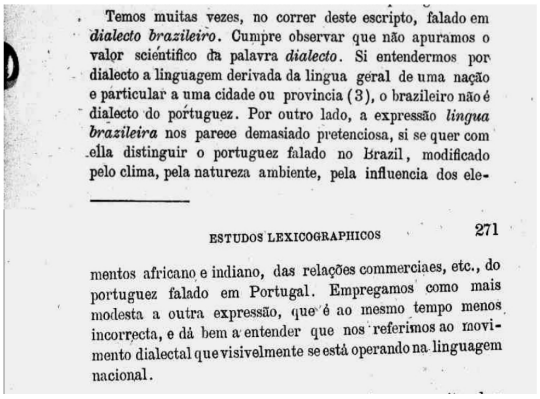
Esse processo consiste na seleção, conversão, correção e limpeza dos documentos que constituem o córpus (PEREIRA; GONÇALVES; SANTOS, 2013). Primeiramente são acessadas as Revistas

Tabela 4.1: Córpus Revista Brazileira.

Documentos Históricos	Autor	Revista Brazileira
Estudos Lexicographicos Do Dialecto Brazileiro IV	Macedo Soares	Ano I, tomo IV, abr-jun 1880, p.243-271.
O Dialecto Brazileiro	Pacheco Junior	Ano II, tomo V, jul-set 1880, p.487-495
Estudos Lexicographicos Do Dialecto Brazileiro VII	Macedo Soares	Ano II, tomo VIII, abr-jun 1881, p.118-126
Uma Questão Glottologica	Pacheco Junior	Ano III, tomo IX, jul-set 1881, p.502-510
Questões de Linguistica	Paranhos da Silva	Ano II, tomo VII, out-dez 1881, p.276-284

Brazileiras disponibilizadas na Hemeroteca Digital Brasileira no formato PDF e, em seguida são convertidas para um documento em formato ODT ou DOC. Essa tarefa é facilitada pois esses documentos já tem seu conteúdo reconhecido pelo OCR, tornando possível converter manualmente os documentos do formato PDF para o formato ODT ou DOC apenas com o uso de ferramentas livremente disponíveis, como um visualizador de documentos PDF e um processador de texto como o *OpenOffice Writer*.

A título de exemplo é apresentado na Figura 4.1 um pequeno trecho das páginas 270 e 271 do documento "Estudos Lexicographicos Do Dialecto Brazileiro IV" de Macedo Soares que faz parte da Revista Brazileira, Ano I, tomo IV, abr-jun 1880, p.243-271. Podem ser observados o resultado da Conversão do documento PDF original (Figura 4.1(a)) para um documento em formato ODT (Figura 4.1(b)).



Temos muitas vezes, no correr deste escripto, falado em *dialecto brazileiro*. Cumpre observar que não apuramos o valor scientifico da palavra *dialecto*. Si entendermos por dialecto a linguagem derivada da lingua geral de uma nação e particular a uma cidade ou provincia (3), o brazileiro não é dialecto do portuguez. Por outro lado, a expressão *lingua brazileira* nos parece demasiado pretenciosa, si se quer com ella distinguir o portuguez falado no Brazil, modificado pelo clima, pela natureza ambiente, pela influencia dos ele-

ESTUDOS LEXICOGRAPHICOS 271

mentos africano e indiano, das relações commerciaes, etc., do portuguez falado em Portugal. Empregamos como mais modesta a outra expressão, que é ao mesmo tempo menos incorrecta, e dá bem a entender que nos referimos ao movimento dialectal que visivelmente se está operando na linguagem nacional.

(a) Documento original em PDF extraído da HDB

Temos muitas vezes, no correr deste escripto, falado em dialecto brazileiro. Cumpre observar que não apurámos o valpr scientifico da palavra dialecto. Si entendermos por dialecto a linguagem derivada da lingua geral de uma nação e particular a uma cidade ou provincia (3), o brazileiro não é dialecto do portuguez. Por outro lado, a expressão lingua brazileira nos parece demasiado pretenciosa, si se quer com ella distinguir o portuguez falado no Brazil, modificado pelo clima, pela natureza ambiente, pela influencia dos ele-

ESTUDOS LEXICOGRAPHICOS 271
mentos africano e indiano, das relações commerciaes, etc, do portuguez falado em Portugal. Empregamos como mais modesta a outra expressão, que'é ao mesmo tempo menos incorrecta, e dá bem a entender que nos
'| referimos ao movimento
dialectal que visivelmente se está operando na-linguagem nacional»

(b) Conversão para ODT ou DOC e Correção dos Erros de OCR

Figura 4.1: Conversão do documento original em formato PDF para o formato ODT ou DOC e Correção dos Erros de OCR.

Após os documentos serem convertidos para o formato ODT, o próximo passo é a revisão

e correção dos erros originados pelo OCR em comparação aos documentos originais. Uma primeira correção é realizada por colaboradores do projeto e uma segunda revisão é feita pelo especialista do sistema, esse passo é importante para que esse tipo de erro não interfira nas futuras análises dos documentos. Os erros mais comuns encontrados são alteração de letras, acentuação, pontuação, junção de palavras, além disso, as linhas com inclinação acentuada não são capturadas pelo OCR e necessitam ser digitadas. Na Figura 4.1(b) são exemplificadas algumas correções efetuadas: "braziieiro" por "brazileiro", "apurámos" por "apuramos", "que'é" por "que é", entre outros. Nesses casos todas as alterações devem ser feitas manualmente.

Com esse processo são obtidos os documentos do córpus de amostra em formato ODT ou DOC corrigido, e as fichas catalográficas contendo os metadados que identificam cada documento e as informações sobre a correção.

4.1.2 Anotação manual de uma amostra do córpus

A partir da análise dos documentos realizada pelo engenheiro de ontologia junto ao especialista de domínio, é possível determinar os termos chave que identificam o domínio que, nesse caso, são os termos Língua, Dialeto e Idioma. Com os termos chave determinados, busca-se as relações no contexto em que esses estão inseridos como, por exemplo, "lingua portugueza" ou "dialecto brazileiro", entre outros. Com isso, é gerado um conjunto de documentos com anotações preliminares (ou *Gold Standard* inicial) que marcam os termos chave e contextos, e uma taxonomia inicial que representa o conhecimento descoberto.

Para apoiar o processo de análise ontológica é realizado um *Brainstorming*, através de um procedimento de anotação manual nos documentos por anotadores do domínio, com o intuito de avaliar a concordância entre essas anotações e o *Gold Standard* inicial e, assim, possibilitar uma análise para verificar se o *Gold Standard* inicial consegue expressar adequadamente o domínio discutido.

No procedimento de anotação é utilizada como referência uma amostragem dos documentos apresentados na Tabela 4.1. Essa amostragem, selecionada pelo especialista de domínio, é formada pelos documentos "O Dialecto Brazileiro" de Pacheco Junior, "Questões de Linguistica" de Paranhos da Silva e "Estudos Lexicographicos Do Dialecto Brazileiro VII" de Macedo Soares, por apresentarem o maior número de ocorrências dos termos e conceitos relacionados ao domínio da Constituição da Língua Portuguesa no Brasil.

Aos anotadores foi fornecido um conjunto de diretrizes para anotação denominado "manual do anotador". O manual indica aos anotadores os passos para a realização da anotação,

facilitando o trabalho e buscando a minimização de erros. As etapas contidas no manual são:

1. Ler todo o documento com atenção e não fazer nenhuma anotação.
2. Ler o documento novamente e identificar os termos relevantes iniciais (Língua, Dialeto e Idioma).
3. Ler o documento pela terceira vez e anotar o contexto onde ocorrem relacionamentos com os termos primários.
4. Anotar no documento algum contexto que seja relacionado com o tema, mesmo que os termos primários não estejam presentes.
5. Descrever uma observação do por quê esse contexto foi marcado.
6. Relatar dúvidas, incertezas e questionamentos quanto às anotações.

Conforme o manual, os anotadores são orientados a anotar inicialmente os termos chave do domínio. Nesse caso, os termos chave são Língua, Dialeto e Idioma. Depois disso deve ser anotado qual é o contexto em que esse termo está inserido, com isso várias adjetivações e relações podem ser identificadas. No próximo passo o anotador deve identificar e marcar termos ou conceitos que sejam relevantes, mesmo que os termos chave não estejam presentes, por exemplo, "Falado no Brasil". E por último, o anotador deve relatar as dificuldades encontradas, as dúvidas e incertezas que surgem durante a anotação.

Uma discussão da teoria que envolve o contexto em torno do domínio foi apresentada aos anotadores, no entanto, eles não foram treinados para as anotações e deveriam seguir unicamente os passos descritos no manual. A escolha de realizar uma anotação deliberadamente grosseira e vagamente definida (DALIANIS; VELUPILLAI, 2010), foi feita para possibilitar a descoberta de como os anotadores compreendem o contexto e determinar a diferença em relação a anotação preliminar, ou seja, as anotações apontadas no *Gold Standard* inicial. Depois disso, os resultados são utilizados para refinar o consenso entre as anotações e possibilitar a geração de uma anotação expressiva do domínio, que permite a extração dos conceitos e relacionamentos para a construção da ontologia, esse consenso é realizado pelo especialista de domínio.

Avaliação das Anotações

As anotações foram feitas por 4 anotadores de domínio, sendo 1 estudante de mestrado e 3 de graduação da área de Letras, identificados por A1, A2, A3 e A4, respectivamente.

Através da medição *Inter-Annotator Agreement* podemos encontrar as inconsistências entre os anotadores e definir o quanto o *Gold Standard* está apropriado e com isso desenvolver o Gold Standard final (VELUPILLAI et al., 2009) (MAKS; VOSSEN, 2010). As métricas Precisão, Revocação e *F-Measure* instrínsecas no campo de Recuperação de Informação (FRAKES; BAEZA-YATES, 1992), podem também ser utilizadas quando queremos comparar o desempenho de cada anotador ao Gold Standard.

Na Figura 4.2 podem ser visualizados os resultados para a métrica F-Measure-Lenient entre o Gold Standard e as anotações de cada anotador. Podemos perceber que quando os termos Dialeto, Idioma, Lingua são observados como termos chave isolados, os resultados de anotação correspondem ao ideal, ou seja, apresentam resultados próximos ou iguais ao Gold Standard(100%). Contudo, ao analisarmos o contexto, isto é, quando ocorrem as relações entre os termos chave ou contextos sobre o domínio sem nenhum termo chave, os números se aproximam de 0.6 para *F-measure Lenient*. Consideramos utilizar a medida Lenient por estar de acordo com o objetivo de analisar a proximidade de anotações entre os pares de anotadores e entre o *Gold Standard*.

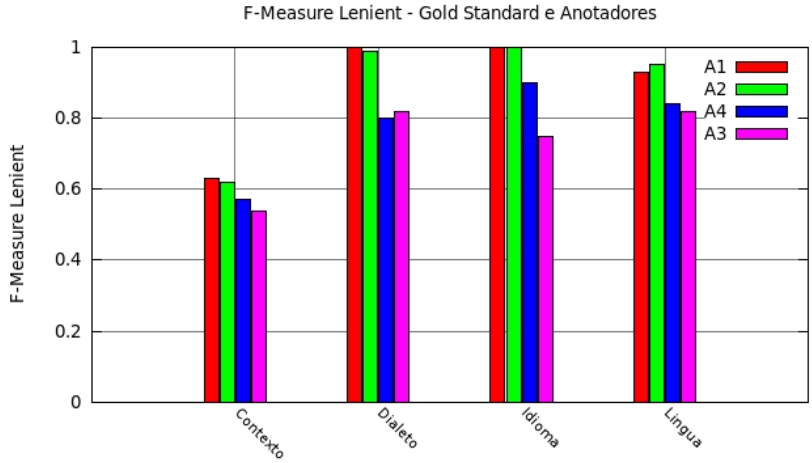


Figura 4.2: Métrica *F-measure Lenient* entre o Gold Standard e as anotações dos anotadores do domínio.

Os resultados encontrados indicam que o *Inter-Annotator Agreement(IAA)* para a anotação "contexto", a de maior interesse, ficou em torno de 60% para o *F1-score*. Os resultados condizem com os resultados encontrados no trabalho de Dalianis e Velupillai (2010), onde ocorre o refinamento de um *Gold Standard* para textos clínicos.

Através disso, o especialista pode determinar com clareza quais os termos, conceitos e relacionamentos que descrevem o domínio da constituição da Língua Portuguesa no Brasil. Com os resultados, percebeu-se que algumas relações encontradas devem ser consideradas por refle-

tirem o domínio discutido como, por exemplo, “lingua portugueza no Brazil”, “dialecto brazileiro”, “brazileiro antigo”, entre outras. Esses resultados auxiliam na construção do consenso entre o *Gold Standard* inicial e as anotações relevantes entre os anotadores, resultando no *Gold Standard* final.

O *Gold Standard* final contemplou as seguintes observações: caracterizando a Espacialidade mostraram-se pertinentes a consideração sobre os povos, o que permite opor de um lado os países lusófonos como Brazil e Portugal (e os derivados brazileiro e portuguez) aos demais países como França, Inglaterra, Castela, Galízia (franceza, ingleza, castelhano, galliziano, entre outras). No que diz respeito à Modalidade, foi importante considerar “falar” e seus derivados, como por exemplo falada, falado, a fim de diferenciar ocorrências em que havia a oposição entre língua falada e língua escrita. Caracterizando a Temporalidade o refinamento do *Gold Standard* refletiu, ainda, a oposição muito recorrente entre “atual” e “antigo” levando em consideração o período em que o texto foi escrito e o período sobre o qual ele falava. Essas são apenas algumas observações, no decorrer desta dissertação outras considerações sobre o domínio são apresentadas.

A partir do *Gold Standard* final, o engenheiro de ontologia implementa a ontologia que expressa o conhecimento adquirido. Lembrando que o processo é iterativo e permite que novas classes, propriedades, instâncias e axiomas, sejam adicionados no decorrer da implementação, caso sejam necessários.

Nesta dissertação denominamos classes e instâncias presentes na ontologia de domínio como **Conceitos Primitivos** (MORAIS; AMBRÓSIO, 2007). A associação entre os conceitos primitivos, os relacionamentos hierárquicos, os relacionamentos de propriedade ou os obtidos por inferência, são denominados **Conceitos Derivados**.

A Ontologia InstrumentoLinguistico gerada a partir dos dados coletados é apresentada e discutida na próxima seção.

4.2 Ontologia InstrumentoLinguistico

Nessa seção são apresentados extratos da Ontologia InstrumentoLinguistico. Nesta ontologia são evidenciados os termos que buscam representar o domínio da Constituição da Língua Portuguesa no Brasil.

Nos extratos da ontologia podem ser visualizados, além das classes, alguns relacionamentos e propriedades. Os relacionamentos hierárquicos estão explícitos pela taxonomia representados

pelas setas indicadas por *é-um (is-a)*. As propriedades de objeto (*Object Property)* podem ser visualizadas pelas setas em azul que relacionam as classes. Esses relacionamentos permitem construir conceitos derivados que enriquecem o conhecimento obtido nos documentos.

A Ontologia InstrumentoLinguistico pode ser visualizada na íntegra com algumas instâncias no Apêndice A. No Apêndice B pode ser visualizada a representação da ontologia em OWL.

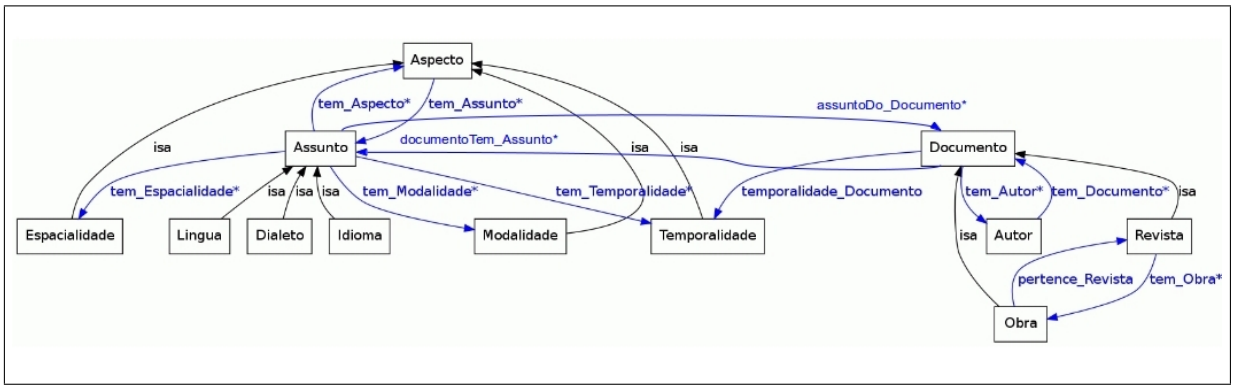


Figura 4.3: Ontologia InstrumentoLinguistico: Assunto, Aspecto e Documento.

Primeiramente, podem ser observadas na Figura 4.3, as classes Assunto e Aspecto que abrangem a maioria das classes que representam o domínio. A classe Assunto contempla os termos chave ou conceitos primitivos (MORAIS; AMBRÓSIO, 2007) que fazem parte do domínio. Esses conceitos primitivos são representados pelas classes Lingua, Dialeto, Idioma, enquanto a classe Outros_Assuntos representa os termos chave palavra, termo e vocábulo. A classe Aspecto é especializada em Espacialidade, Modalidade e Temporalidade, que refletem os aspectos analisados nos documentos.

Em seguida, pode ser observada a classe Documento que é especializada pelas classes Revista e Obra, e a classe Autor. Por exemplo, para identificar que uma Obra pertence a uma Revista e tem um Autor, de acordo com a ontologia:

Obra *pertence_Revista* **Revista** *é-um* **Documento** *tem_Autor* **Autor**

É importante a consideração dos documentos na ontologia pois a discussão sobre a Língua Portuguesa no Brasil se deu por meio de documentos e periódicos históricos. Esses documentos revelam posições distintas de brasileiros e portugueses em relação à constituição da Língua Portuguesa no Brasil. Além disso, essas informações são imprescindíveis para a correta identificação de um documento.

Ainda considerando a classe Documento, tem-se as propriedades inversas do relacionamento com a classe Assunto. Um Documento tem a propriedade *documentoTem_Assunto* As-

sunto. E a classe Assunto tem a propriedade *assuntoDo_Documento* Documento (ver Figura 4.4).

```
<owl:ObjectProperty rdf:about="#documentoTem_Assunto">
    <rdfs:domain rdf:resource="#Documento"/>
    <owl:inverseOf rdf:resource="#assuntoDo_Documento"/>
    <rdfs:range rdf:resource="#Assunto"/>
  </owl:ObjectProperty>
```

Figura 4.4: Representação em OWL da Ontologia InstrumentoLinguistico: Propriedades Inversas *documentoTem_Assunto* e *assuntoDo_Documento*.

Na Figura 4.5 podem ser visualizadas as especializações da classe Espacialidade pelos conceitos primitivos representados pelas classes Cidade, Pais, Regiao, Estado e Provincia, que referenciam um lugar. A classe Pais é especializada nas classes disjuntas Paises_Lusofonos e Outros_Paises (ver Figura 4.6).

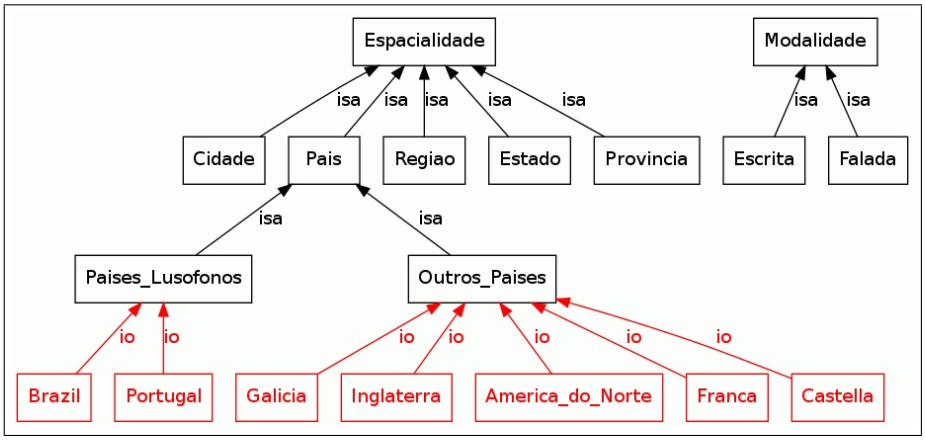


Figura 4.5: Ontologia InstrumentoLinguistico: Espacialidade com algumas instâncias de Paises_Lusofonos e Outros_Paises e Modalidade.

A classe Paises_Lusofonos pode conter as instâncias Brazil e Portugal, e na classe Outros_Paises podem fazer parte Galícia, Inglaterra, America do Norte, França e Castella. Esses são alguns exemplos de países lusófonos e países não lusófonos, sendo que no mundo real existem inúmeros outros casos que não são discutidos neste trabalho.

```
<owl:Class rdf:ID="Paises_Lusofonos">
  <owl:disjointWith>
    <owl:Class rdf:ID="Outros_Paises"/>
  </owl:disjointWith>
  <rdfs:subClassOf rdf:resource="#Pais"/>
</owl:Class>
```

Figura 4.6: Representação em OWL da Ontologia InstrumentoLinguistico: Classes Disjuntas Paises_Lusofonos e Outros_Paises.

```
<owl:Class rdf:ID="Falada">
  <owl:disjointWith>
    <owl:Class rdf:ID="Escrita"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="Modalidade"/>
  </rdfs:subClassOf>
</owl:Class>
```

Figura 4.7: Representação em OWL da Ontologia InstrumentoLinguistico: Classes Disjuntas Falada e Escrita.

Em Modalidade são representados os conceitos primitivos pelas classes disjuntas falada e escrita. São disjuntas pois se referem a modos opostos, ou a modalidade é falada ou é escrita (ver Figura 4.7). Essas ocorrências de modo podem existir em um relacionamento da classe Assunto com a classe Modalidade representado pela propriedade *tem_Modalidade*. Por exemplo, para expressar o conceito derivado "língua falada" conforme a ontologia, tem-se:

Lingua *tem_Modalidade* **Falada**

Como pode ser observado na Figura 4.8, Lingua *é-um* Assunto e Falada *é-um* Modalidade. Sendo assim, o relacionamento *tem_Modalidade* persiste para a classe especializada.

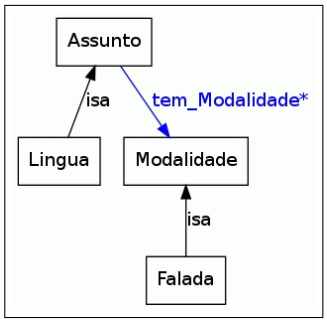


Figura 4.8: Ontologia InstrumentoLinguistico: Lingua *tem_Modalidade* Falada.

A Temporalidade, conforme Figura 4.9, representa as classes que configuram tempo. Essa classe é especializada nas classes Temporalidade_Precisa e Temporalidade_Imprecisa.

A classe Temporalidade_Precisa tem como propriedade tipo de dado *(DataType Property)* século, década, ano, mês e data, que representam uma temporalidade exata (ver exemplos na Figura 4.10). Por exemplo, o conceito derivado "ano de publicação de um Documento" pode ser expresso por um relacionamento *temporalidade_Documento* entre a classe Documento e Temporalidade. Conforme a ontologia, tem-se:

Documento *tem_Temporalidade* **Temporalidade_Precisa : ano.**

Como pode ser observado, o documento *tem_Temporalidade* Temporalidade_Precisa e o tipo de dado de interesse nessa classe é ano.

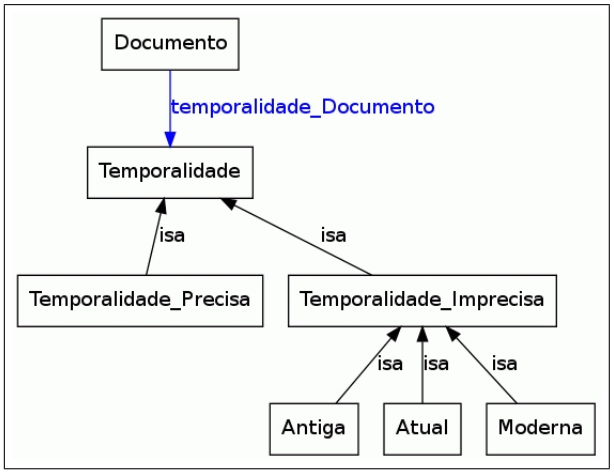


Figura 4.9: Ontologia InstrumentoLinguistico: Temporalidade.

```
<owl:DatatypeProperty rdf:ID="seculo">
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
  <rdfs:domain rdf:resource="#Temporalidade_Precisa"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="decada">
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
  <rdfs:domain rdf:resource="#Temporalidade_Precisa"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="ano">
  <rdfs:domain rdf:resource="#Temporalidade_Precisa"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#int"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="mes">
  <rdfs:domain rdf:resource="#Temporalidade_Precisa"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="data">
  <rdfs:domain rdf:resource="#Temporalidade_Precisa"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#date"/>
</owl:DatatypeProperty>
```

Figura 4.10: Representação em OWL da Ontologia InstrumentoLinguistico: Temporalidade_Precisa e seus Tipos de Dados.

A classe Temporalidade_Imprecisa contempla conceitos primitivos de tempo que não são exatos, como as classes Atual, Antiga e Moderna. Essa classe permite identificar no documento relações de tempo como, por exemplo, o conceito derivado “língua actual”. Nesse caso, de acordo com a ontologia:

Lingua *tem_Temporalidade* **Atual**

Como pode ser observado na Figura 4.11, Lingua *é-um* Assunto e Atual *é-um* Temporalidade_Imprecisa que *é-um* Temporalidade. Sendo assim, o relacionamento *tem_Temporalidade* persiste para as classes especializadas.

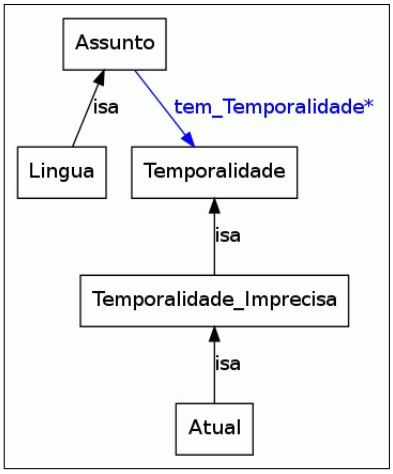


Figura 4.11: Ontologia InstrumentoLinguistico: Lingua *tem_Temporalidade* Atual.

A classe Povos, Figura 4.13, representa a naturalidade de um país, ora considerando realmente um povo, ora considerando o falar de um povo. Essa distinção é percebida quando uma instância de povo está associada com uma instância da classe Assunto, então, nesse caso, é o falar de um povo.

A classe Povos está relacionada com a classe Pais por duas propriedades que são inversas. Dessa forma, a classe Povos é *natural_de* Pais, e a classe Pais *tem_Naturalidade* Povos (ver Figura 4.12).

```
<owl:ObjectProperty rdf:ID="natural_de">
  <owl:inverseOf>
    <owl:ObjectProperty rdf:ID="tem_Naturalidade"/>
  </owl:inverseOf>
  <rdfs:domain rdf:resource="#Povos"/>
  <rdfs:range rdf:resource="#Pais"/>
</owl:ObjectProperty>
```

Figura 4.12: Representação em OWL da Ontologia InstrumentoLinguistico: propriedades inversas *natural_de* e *tem_Naturalidade*.

Por exemplo, para expressar o conceito derivado “Dialecto brazileiro”, de acordo com a ontologia:

Dialeto *tem_Naturalidade* brazileiro

Conforme a Figura 4.13, a instância dialecto faz parte da classe Dialeto que *é-um* Assunto que *tem_Espacialidade* Espacialidade. Pais *é-um* Espacialidade e Pais *tem_Naturalidade* Povos. A instância “brazileiro” é *natural_de* Brazil que é uma instância de Paises_Lusofonos que *é-um* Pais. Pela transitividade pode-se inferir que Dialeto *tem_Naturalidade* brazileiro.

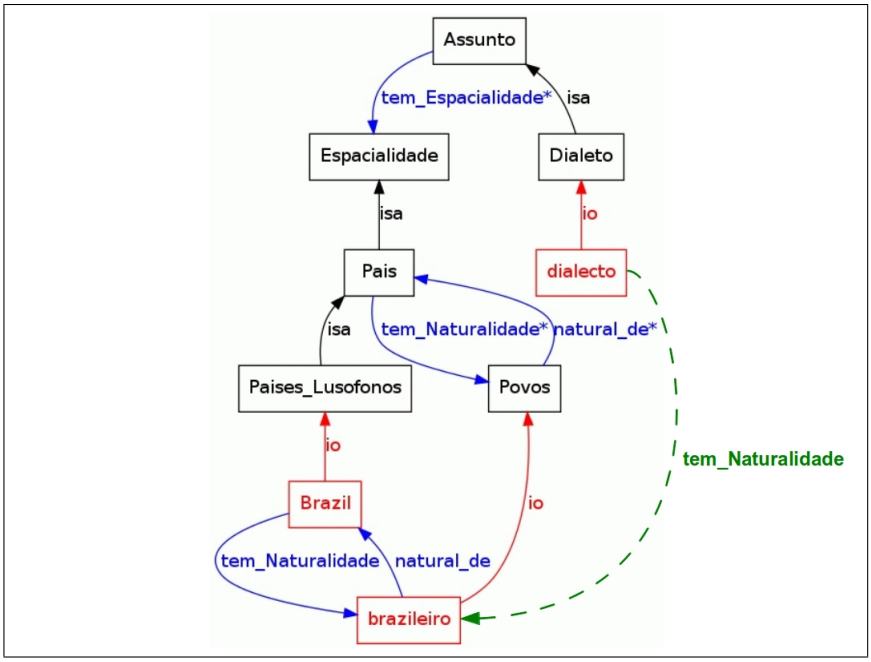


Figura 4.13: Ontologia InstrumentoLinguistico: Povos.

Por último, a classe Qualificação especifica a qualificação da classe Assunto. Essa relação pode ser observada na Figura 4.14. Alguns exemplos de instâncias de qualificação são: corrente, popular, litteraria, juridica, sagrada.

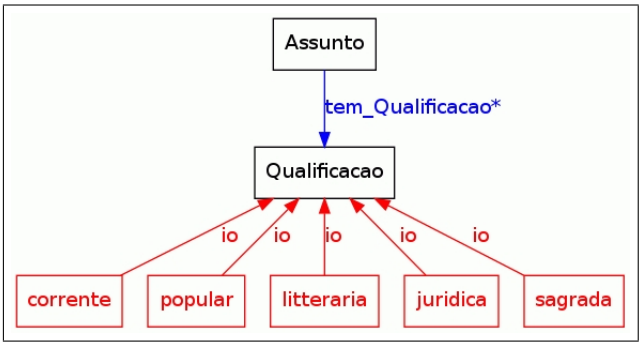


Figura 4.14: Ontologia InstrumentoLinguistico: classe Qualificacao com algumas instâncias.

Essa propriedade permite relacionar a classe Assunto e Qualificacao com adjetivações encontradas. Por exemplo, para expressar o conceito derivado "língua corrente", de acordo com a ontologia, tem-se:

Lingua *tem_Qualificacao* **Qualificacao : corrente**

É importante mencionar que na ontologia as classes, propriedades e instâncias estão sem acentuação para evitar eventual incompatibilidade com *software*.

Capítulo 5

ANOTAÇÃO SEMÂNTICA AUTOMÁTICA BASEADA EM ONTOLOGIA

5.1 Arquitetura da Abordagem para Anotação Semântica

A abordagem de Anotação Semântica Automática Baseada em Ontologia *(Automatic Ontology-based Semantic Annotation Approach)* tem por objetivo anotar semanticamente documentos de acordo com uma ontologia de domínio.

A arquitetura da abordagem para anotação semântica em documentos pode ser visualizada na Figura 5.1. A ideia principal é transformar os documentos para permitir a extração de informação e, assim, possibilitar a anotação semântica nos documentos. Para isso, a abordagem é dividida em duas etapas: o Pré-Processamento e a Extração de Informação Baseada em Ontologia.

A representação foi baseada na técnica de Ross (1977), *Structured Analysis and Design Technique (SADT)*. Nesse diagrama, de acordo com a caixa-legenda, os retângulos representam os processos da abordagem. As setas que entram pelo lado esquerdo dos retângulos representam as entradas de dados, e as setas que saem pelo lado direito representam as saídas geradas em cada etapa. As setas superiores representam os controles que orientam a execução de cada etapa. Já as setas inferiores representam os mecanismos, os participantes e as ferramentas que auxiliam ou automatizam a execução das etapas.

Como pode ser observado na Figura 5.1, a entrada de dados é composta por um córpus contendo documentos que necessitam ser anotados semanticamente. Esse córpus pode conter documentos em diversos formatos, como PDF, XML, HTML, MS Word e TXT. Os documentos passam pela etapa de pré-processamento que é automática e utiliza a biblioteca OpenNLP para

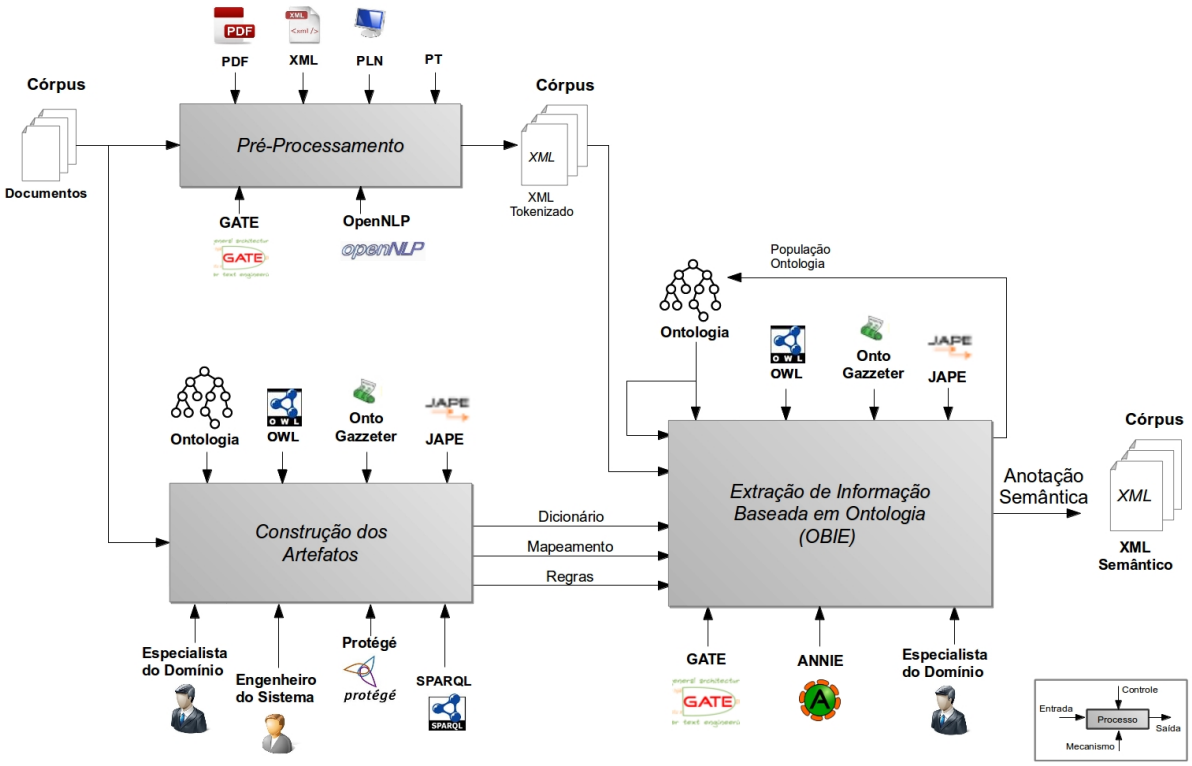


Figura 5.1: Arquitetura da Abordagem para Anotação Semântica.

as tarefas de Processamento de Língua Natural (PLN).

Paralelamente, deve ser realizada a etapa de Construção dos Artefatos. Nessa etapa, o engenheiro do sistema com a colaboração do especialista de domínio, desenvolvem os artefatos que serão utilizados na próxima etapa. Os recursos ANNIE *(A Nearly-New Information Extraction System)*, JAPE *(Java Annotation Patterns Engine)* e *OntoGazzeter* apoiam o processo. O resultado é um Dicionário de Conceitos Primitivos e um Mapeamento desse Dicionário com a ontologia de domínio, além de um conjunto de Regras que implementa a anotação dos conceitos derivados.

Em seguida, os artefatos, os documentos pré-processados e transformados em documentos XML Tokenizado, e a ontologia de domínio são entradas para a etapa de extração de informação baseada em ontologia (OBIE). A ontologia, além de ser uma entrada para a etapa, também constitui o controle do processo de extração, pois conduz a implementação das tarefas dessa etapa.

A saída é um córpus contendo documentos no formato XML. O XML semântico, assim denominado neste trabalho, contém marcações que representam os conceitos primitivos e os

conceitos derivados extraídos da ontologia de domínio. A anotação semântica obtida proporciona identificar o conhecimento contido nos documentos.

A construção do protótipo que implementa a abordagem foi inteiramente baseada no *framework* GATE.

5.2 Pré-Processamento

O Pré-Processamento é a etapa que tem por objetivo preparar os documentos para permitir que seja realizado corretamente o processo de extração de informação. Na Figura 5.2 está representada a arquitetura do pré-processamento da abordagem discutida neste trabalho.

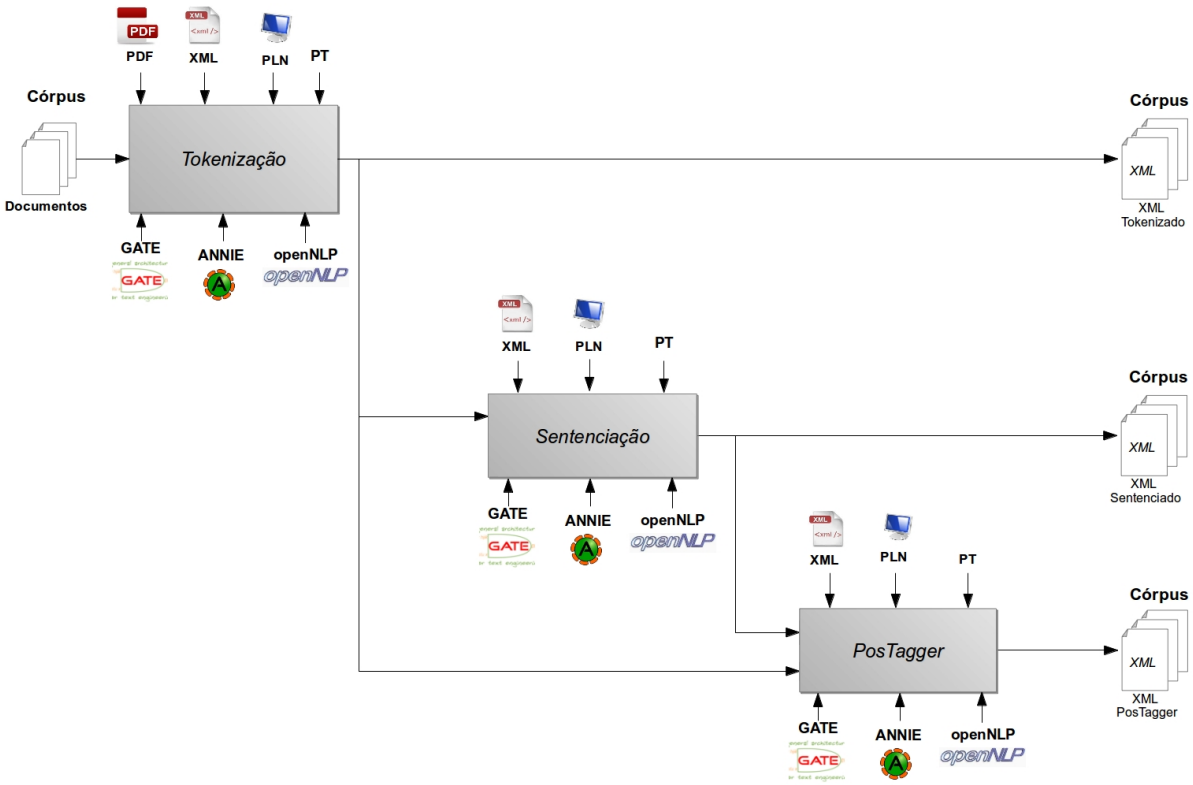


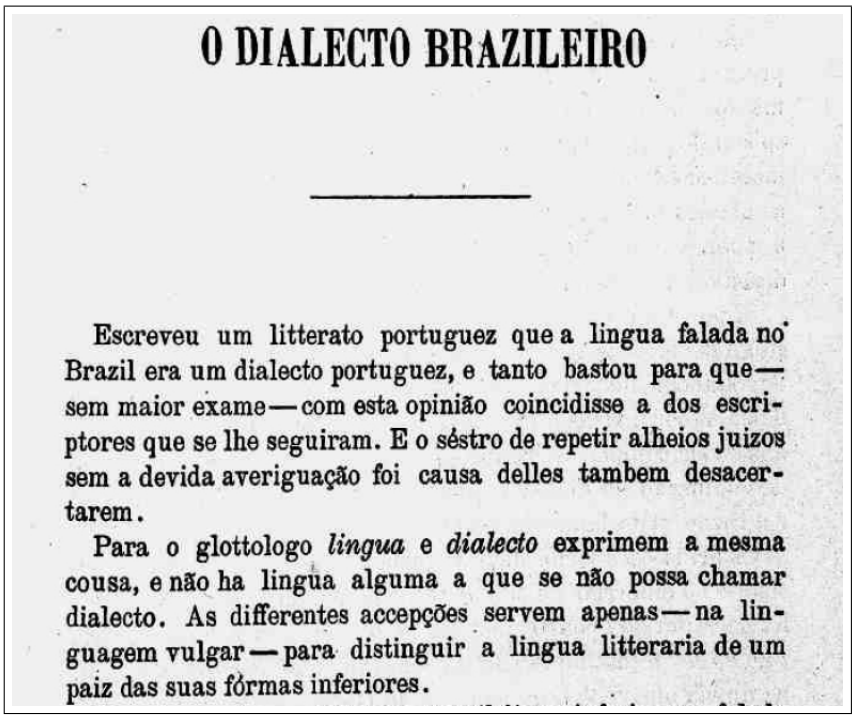
Figura 5.2: Arquitetura do Pré-Processamento.

Como pode ser observado, um córpus é a entrada para a etapa do pré-processamento que realiza a Tokenização. O objeto de estudo deste trabalho é um córpus que contém documentos no formato PDF, conforme exemplo apresentado na Figura 5.3.

A etapa de Tokenização é um processo imprescindível e o único necessário de ser realizado no pré-processamento deste trabalho. O resultado dessa etapa, são os documentos no formato

XML com as marcações que identificam o tipo *Token* e o tipo *SpaceToken* para os espaços em branco. Um exemplo do documento XML Tokenizado pode ser observado na Figura 5.4.

Todas as etapas relacionadas ao pré-processamento podem utilizar a biblioteca OpenNLP para a Língua Portuguesa atual, devido ao fato que os documentos históricos discutidos neste trabalho fazem parte de um Português do final do século XIX, momento onde, as características que determinam um *token* ou sentença são iguais as atuais.



O DIALECTO BRAZILEIRO

Escreveu um litterato portuguez que a lingua falada no Brazil era um dialecto portuguez, e tanto bastou para que— sem maior exame—com esta opinião coincidisse a dos escriptores que se lhe seguiram. E o séstro de repetir alheios juizos sem a devida averiguação foi causa delles tambem desacertarem.

Para o glottologo *lingua* e *dialecto* exprimem a mesma cousa, e não ha lingua alguma a que se não possa chamar dialecto. As differentes accepções servem apenas—na linguagem vulgar—para distinguir a lingua litteraria de um paiz das suas fórmas inferiores.

Figura 5.3: Trecho da Revista Brazileira: documento "O Dialecto Brazileiro".

O_DIALECTO_BRAZILEIRO_Token.xml

```
1 <Token>O</Token> <Token>DIALECTO</Token> <Token>BRAZILEIRO</Token>
2 <Token>Escreveu</Token> <Token>um</Token> <Token>litterato</Token> <Token>portuguez</Token>
  <Token>que</Token> <Token>a</Token> <Token>lingua</Token> <Token>falada</Token> <Token>no</
  Token> <Token>Brazil</Token> <Token>era</Token> <Token>um</Token> <Token>dialecto</Token>
  <Token>portuguez</Token><Token>,</Token> <Token>e</Token> <Token>tanto</Token> <Token>bastou</
  Token> <Token>para</Token> <Token>que</Token> <Token>—</Token> <Token>sem</Token> <Token>maior</
  Token> <Token>exame</Token> <Token>—</Token> <Token>com</Token> <Token>esta</Token>
  <Token>opinião</Token> <Token>coincidisse</Token> <Token>a</Token> <Token>dos</Token>
  <Token>escriptores</Token> <Token>que</Token> <Token>se</Token> <Token>lhe</Token>
  <Token>seguiram</Token><Token>.</Token> <Token>E</Token> <Token>o</Token> <Token>séstro</Token>
  <Token>de</Token> <Token>repetir</Token> <Token>alheios</Token> <Token>juizos</Token>
  <Token>sem</Token> <Token>a</Token> <Token>devida</Token> <Token>averiguação</Token>
  <Token>foi</Token> <Token>causa</Token> <Token>delles</Token> <Token>também</Token>
  <Token>desacertarem</Token><Token>.</Token>
```

Figura 5.4: Documento XML com as marcações do tipo *Token* no documento " O Dialecto Brazileiro".

Neste trabalho, foi utilizada apenas a etapa de Tokenização. Adicionalmente, podem ser realizadas as etapas de Sentenciação e de anotação morfossintática *(PosTagger)*, caso alguma análise particular necessite desses processos. Na nossa abordagem essas etapas não estão sendo

realizadas e mostramos que são dispensáveis para a realização da anotação semântica dos documentos.

Essas etapas não estão sendo consideradas neste trabalho, pois não é necessário realizar análise morfossintática nos documentos. Porém, elas estão representadas na arquitetura apresentada na Figura 5.2 para indicar em que fase essa análise pode ser adicionada, caso seja necessário.

As etapas de Sentenciação e de *PosTagger* são etapas alternativas, porém, caso sejam realizadas, o arquivo XML Tokenizado pode ser utilizado como entrada. O resultado dessa etapa é um arquivo no formato XML, o XML Sentenciado, com as marcações que identificam os *tokens*, as sentenças e os *splits* nos documentos.

O documento XML Sentenciado é o arquivo de entrada da etapa de *PosTagger*, que é possível ser realizado utilizando a biblioteca OpenNLP, mesmo que em documentos históricos como nessa abordagem. E o resultado é um arquivo no formato XML que possui as marcações que identificam as categorias morfossintáticas de cada token.

No entanto, nessa etapa, é necessária a intervenção de um linguísta para determinar se essas anotações estão corretas. Existe a chance de ocorrência de alguns erros, pois essa biblioteca foi desenvolvida para tratar de textos escritos em português atual. Nesse sentido, a ideia é que o linguista ou o especialista corrija apenas os trechos no documento em que exista alguma anotação do tipo conceito primitivo ou conceito derivado. Isso reduz a quantidade e o tempo de correção.

Outra alternativa, que está além do escopo deste trabalho, é adicionar uma biblioteca específica para a análise morfossintática em documentos históricos. Os trabalhos descritos no Capítulo 3, que foram desenvolvidos para o tratamento dos documentos históricos pertencentes ao Corpus Tycho Brahe e ao Dicionário Histórico do Português do Brasil (DHPB), podem colaborar com o enriquecimento dessa análise.

5.3 Construção dos Artefatos

O processo automático de anotação semântica baseada em ontologia exige que seja realizada, primeiramente, a Construção dos Artefatos que apoiam o processo de anotação. Na Figura 5.5, pode ser observada a arquitetura que representa o processo de Construção dos Artefatos.

Como pode ser observado, a entrada para a primeira etapa é o córpus contendo os documentos a serem anotados semanticamente. Essa etapa, de Definição dos Requisitos de Anotação,

é realizada exclusivamente pelo especialista de domínio, pois é o especialista que detém as informações pertinentes do que realmente é relevante e precisa ser anotado. Como resultado dessa etapa, um conjunto de requisitos para a anotação é gerado.

Esse conjunto de requisitos é a entrada da próxima etapa, onde ocorre a construção do dicionário. Essa etapa não é trivial, e é discutida na próxima seção. Como resultado dessa etapa, tem-se o Mapeamento do Dicionário e o Mapeamento das Listas do Dicionário para os Conceitos presentes na Ontologia, ambos são apresentados nas próximas seções. Além disso, é realizado um refinamento do conjunto de requisitos para a determinação dos conceitos derivados que devem ser anotados nos documentos.

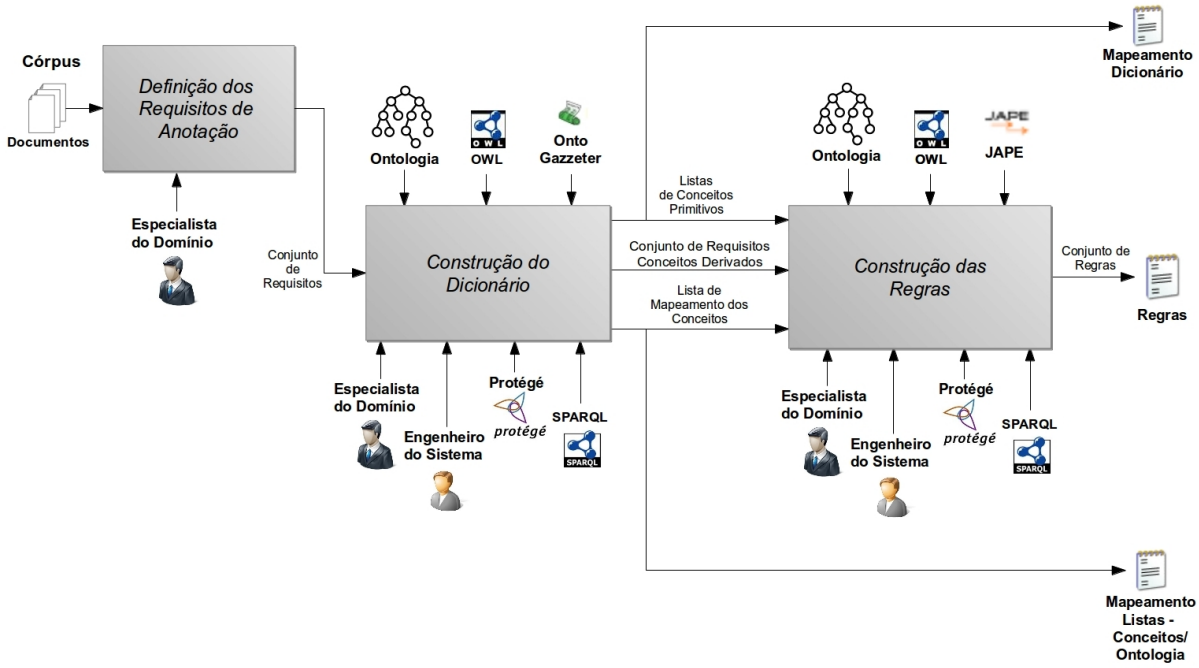


Figura 5.5: Arquitetura da Construção dos Artefatos.

Todas as saídas dessa etapa alimentam a próxima, a etapa de construção das regras. Etapa fundamental para uma adequada anotação semântica nos documentos de interesse. Nessa etapa são definidos os conceitos derivados a partir dos conceitos primitivos já estabelecidos na etapa anterior.

O engenheiro do sistema e o especialista de domínio são os responsáveis por essa etapa. Auxiliados pela ontologia de domínio, e com o apoio na gramática do JAPE, são estabelecidas as regras que conduzem o processo de anotação semântica. Como resultado, é gerado um conjunto de regras para ser aplicado na etapa da anotação semântica.

5.3.1 Construção do Dicionário

A construção do dicionário requer esforço, tanto do Engenheiro do Sistema, quanto do Especialista do Domínio, para obtenção de recursos adequados, com o objetivo de atingir o sucesso no processo de anotação semântica. A arquitetura do processo de Construção do Dicionário, pode ser observada na Figura 5.6.

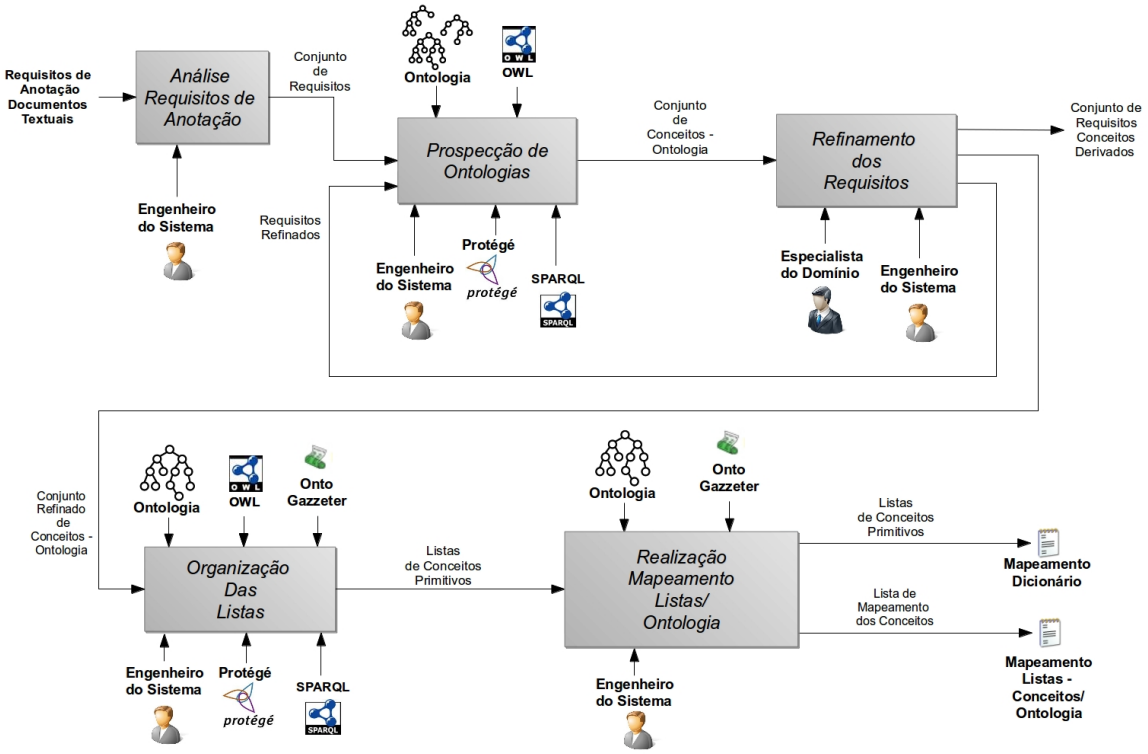


Figura 5.6: Arquitetura da Construção do Dicionário.

Análise dos Requisitos de Anotação

Primeiramente, como entrada no processo, tem-se os Requisitos de Anotação dos Documentos Textuais, estabelecidos anteriormente, pelo especialista. Por exemplo, no objeto de estudo apresentado neste trabalho, o especialista de domínio apresentou sua necessidade de anotação conforme extrato apresentado na Figura 5.7.

Esses requisitos são a entrada da etapa de Análise dos Requisitos de Anotação. Nessa etapa, o engenheiro do sistema faz a análise do requisitos estabelecidos pelo especialista, e gera um conjunto de requisitos analisados que serve como entrada para a próxima etapa, a Prospecção de Ontologias.

Objetivo: Identificar /extrair informação baseada no conteúdo dos periódicos publicados no Brasil no final do século XIX, que discutem a constituição da Língua Portuguesa no Brasil.

Documentos a serem anotados: Um conjunto de documentos digitalizados concernentes à Revista Brazileira, em sua fase que compreende o período de 1879 a 1900, em que, é predominante a discussão da diferença da Língua Portuguesa no Brasil em relação a Portugal.

Termos Chave: Língua, Dialeto e Idioma.

Exemplos de anotação: "língua falada no Brasil", "dialecto brazileiro", "língua brazileira", "portuguez falado no Brazil", "portuguez de Portugal", "dialeto portuguez", "falar de Portugal".

Figura 5.7: Extrato do documento de Requisitos produzido pelo Especialista de Domínio.

Ciclo de Prospecção de Ontologias e Refinamento dos Requisitos

Na etapa de Prospecção de Ontologias, o engenheiro do sistema com o apoio da ferramenta Protégé e da linguagem de consulta SPARQL, investiga nas ontologias de interesse, conceitos primitivos que atendam aos requisitos impostos pelo especialista. Nessa etapa, mais de uma ontologia pode ser utilizada como recurso para atender a demanda necessária de informação pertinente para realização da anotação semântica. No caso de estudo apresentado nesta dissertação, apenas uma ontologia está sendo utilizada, porém, a abordagem suporta a utilização de múltiplas ontologias que devem ser selecionadas pelo engenheiro de sistemas com o aval do especialista de domínio.

Ainda na discussão dessa etapa, é importante considerar que se nenhuma ontologia existente atende aos requisitos impostos, é necessário o desenvolvimento da ontologia que atenda aos requisitos em um processo anterior. Para a realização de nosso estudo de caso, foi necessária a construção de uma ontologia adequada para o propósito de anotar semanticamente documentos históricos do século XIX, que discutem a constituição da língua portuguesa no Brasil. O processo de construção dessa ontologia é discutido no capítulo 4 desta dissertação.

A saída dessa etapa, é um conjunto de conceitos pertencentes às ontologias de domínio em questão. Na Figura 5.8 está representada a abstração de um requisito que forma o conjunto de Conceitos-Ontologia.

Conceito Anotado	Ontologia Utilizada	Conceito da Ontologia

Figura 5.8: Abstração de um requisito Conceito-Ontologia.

Esses conceitos pré-definidos são entrada para o processo de Refinamento dos Requisitos, em que o especialista e o engenheiro discutem se a(s) ontologia(s) escolhidas atendem aos requisitos, ou se é necessário refinar os requisitos e buscar novas bases de conhecimento - ontologias, que atendam o esperado, ou seja, retornando para a etapa anterior Prospecção de

Ontologias. O refinamento dos requisitos é um processo iterativo, que exige análise constante do engenheiro e do especialista, até que cheguem a um consenso e delimitação do escopo de cada ontologia quanto aos conceitos que devem ser utilizados na anotação semântica.

Por exemplo, um requisito do objeto de estudo, é apresentado na Figura 5.9.

Conceito: variação de grafia de Dialeto	*URI*: http://www.owl-ontologies.com/ Instrumento-Linguistico.owl	Domínio de Propriedade de Instância: grafia_dialeto

Figura 5.9: Requisito de variação de grafia de "Dialeto".

A saída dessa etapa é um conjunto de Conceitos-Ontologia Refinados pertencentes à(s) Ontologia(s) e, será utilizado na formação das listas que compõem o dicionário. Além disso, são estabelecidas as diretrizes que formam o conjunto de requisitos para os conceitos derivados, que serão utilizados no processo de Construção de Regras.

Organização das Listas

A etapa de Organização das Listas é alimentada pelos conceitos de interesse pertencentes à ontologia de domínio. Essa etapa tem como controle o recurso ANNIE. O engenheiro, com apoio da ferramenta Protégé e da linguagem SPARQL, constrói e organiza as listas de conceitos que formam o dicionário. Essas listas são formadas pelos conceitos primitivos encontrados na(s) ontologia(s).

As listas são construídas pelo Engenheiro do Sistema, em acordo com os requisitos estabelecidos pelo Especialista de Domínio. Essas listas devem descrever adequadamente os conceitos primitivos, pois influenciam diretamente na anotação semântica dos conceitos derivados.

Para conduzir a tarefa de construção das listas que formam o dicionário, podem ser realizadas consultas *SPARQL* sobre a ontologia de domínio. Essas consultas podem ser executadas na ferramenta Protégé no painel *SPARQL Query*.

O especialista de domínio indicou que os conceitos primitivos Dialeto, Idioma e Lingua, são considerados os termos chave. Pensando nisso, são construídas listas que representam cada um desses termos e as suas variações. O que permite que as anotações sejam feitas em todas as ocorrências do termo, independente da maneira como foi escrito.

Por exemplo, para criar uma lista para o conceito primitivo "dialeto" e suas possíveis variantes de grafia, pode-se executar a consulta, ilustrada na Figura 5.10. Com esse resultado, é possível construir a lista *"dialeto.lst"*, que é composta por todas as ocorrências retornadas pela consulta.

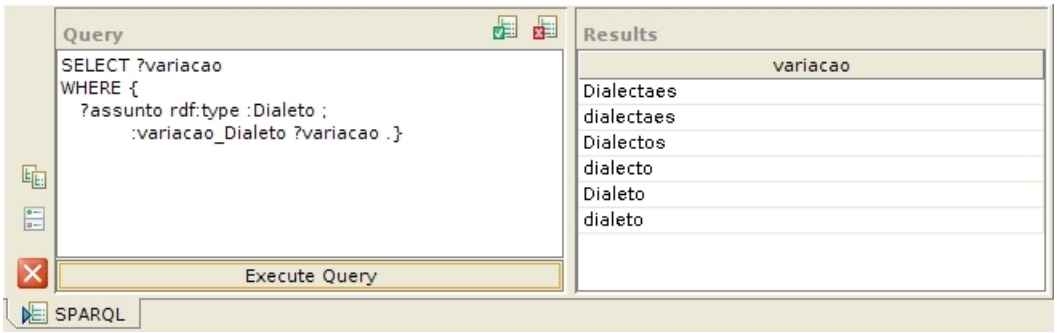


Figura 5.10: Consulta SPARQL: variações do conceito primitivo Dialeto.

Para auxiliar a consulta *SPARQL*, é criada uma instância de Dialeto "grafia_dialeto" que tem como valores para a propriedade tipo de dado *variacao_Dialeto*, as variações de grafia do termo chave em questão.

Como pode ser observado na Figura 5.10, a variável *?variacao* recebe os resultados da seleção entre as instâncias da classe Dialeto, dos valores para a propriedade *variacao_Dialeto*. Esses resultados formam a lista *"dialeto.lst"*.

Outro exemplo, considerando que deve ser anotada a qualificação do dialeto, informando se é corrente, popular, etc, é preciso que seja efetuada uma consulta à classe Qualificacao, conforme pode ser visualizado na Figura 5.11. Nessa consulta, a variável *?qualificacao* seleciona todas as instâncias da classe Qualificacao. Com esse resultado pode ser formada a lista *"qualificacao.lst"*.

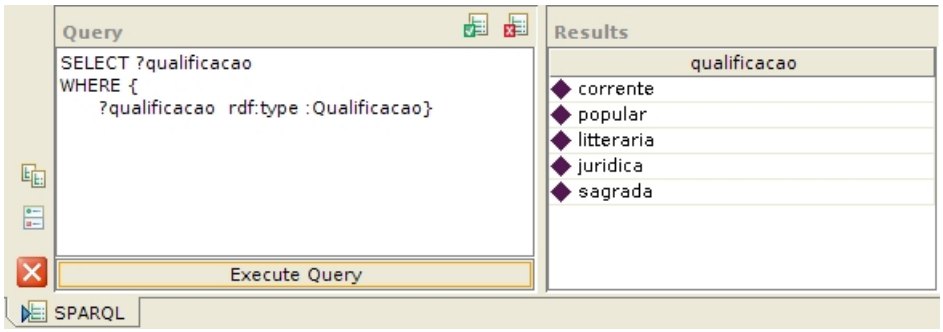


Figura 5.11: Consulta SPARQL: instâncias da classe Qualificacao.

As listas que formam o dicionário são nomeadas com o nome desejado e a extensão *.lst*, por exemplo, *"nome.lst"*. O ideal é que esse nome seja intuitivo e represente adequadamente o conteúdo da lista, todas essas listas devem estar em uma única pasta de arquivos.

Realização do Mapeamento

Em seguida, na Realização do Mapeamento, o conjunto de Listas é mapeado para formar

o Dicionário e é mapeado para a Ontologia de Domínio. Para isso, são utilizadas listas de mapeamento como estabelecido pelo recurso ANNIE.

O arquivo, usualmente no GATE, denominado *"lists.def"*, está representado na Figura 5.12. É um arquivo de mapeamento utilizado para acessar as listas que compõem o dicionário. No arquivo, estão contidos os mapeamentos para todas as listas de interesse. Cada linha representa uma lista individualmente e não existem linhas em branco após a última lista no arquivo.

Esse arquivo deve estar na mesma pasta em que estão as outras listas. O caminho do seu endereço é utilizado no protótipo da abordagem para identificar quais são as listas que formam o dicionário.

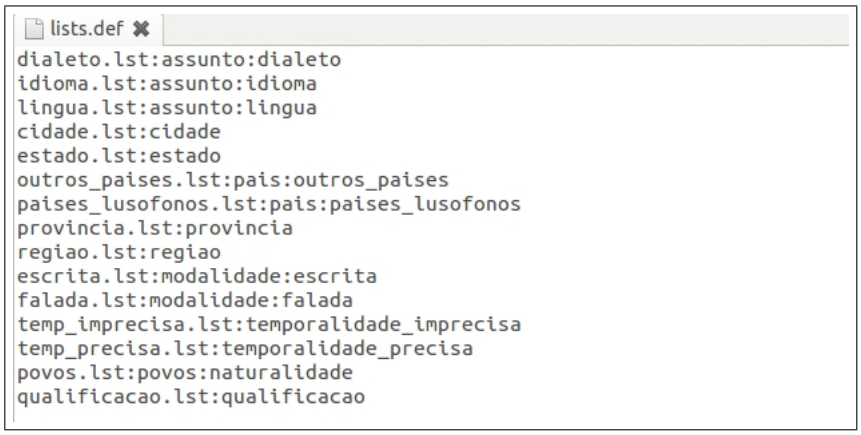


Figura 5.12: Arquivo: *lists.def*.

Além disso, como pode ser observado na Figura 5.12, o mapeamento de uma lista deve ter um maior tipo *(major type)* e, opcionalmente, um menor tipo *(minor type)*. No objeto de estudo desta abordagem, o maior tipo é a classe mais expressiva na ontologia que possui o conceito representado pela lista, e o menor tipo é a informação mais próxima do conceito na Ontologia. Essa informação pode ser uma classe, instância, propriedade de instância ou domínio de propriedade de instância.

Nesse sentido, um exemplo da abstração da representação do mapeamento pode ser observada na Figura 5.13. Onde, a primeira coluna refere-se ao nome da lista, a segunda coluna ao maior tipo, e a terceira ao menor tipo.

Nome.lst	Maior Tipo	Menor Tipo

Figura 5.13: Abstração do mapeamento para formar o dicionário.

O exemplo, de acordo com a primeira linha representada no arquivo *lists.def*, pode ser visualizado na Figura 5.14. Nesse caso, o nome da lista é *"dialeto.lst"*, o maior tipo é a classe *"Assunto"*, e o menor tipo é o domínio de uma propriedade da instância *"grafia_dialeto"*.

dialeto.lst	Assunto	grafia_dialeto

Figura 5.14: Mapeamento da lista: dialeto.lst, Maior Tipo: Assunto, Menor tipo: grafia_dialeto.

Além do mapeamento para o dicionário, existe o mapeamento de cada lista para alguma classe na ontologia. Nesse mapeamento, o menor tipo indica qual é a classe na ontologia que a lista pertence. Considerar a classe que contém a instância, quando o menor tipo for uma instância, propriedade de instância ou domínio de propriedade de instância. Caso não exista menor tipo, considerar a classe representada pelo maior tipo.

O Mapeamento Listas- Conceitos/Ontologia é representado por um arquivo denominado de *"mapping.def"*, que pode ser visualizado na Figura 5.15. Esse arquivo contém as informações para indicar o mapeamento de cada lista para a classe correspondente na ontologia. Cada linha representa um mapeamento individual e não existem linhas em branco após o último item.

mapping.def

```
dialeto.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Dialeto
idioma.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Idioma
lingua.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Lingua
cidade.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Cidade
estado.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Estado
outros_paises.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Outros_Paises
paises_lusofonos.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Paises_Lusofonos
provincia.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Provincia
regiao.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Regiao
escrita.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Escrita
falada.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Falada
temp_imprecisa.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Temporalidade_Imprecisa
temp_precisa.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Temporalidade_Precisa
povos.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Povos
qualificacao.lst:http://www.owl-ontologies.com/InstrumentoLinguistico.owl:Qualificacao
```

Figura 5.15: Arquivo: ***mapping.def*****.**

Pensando nisso, um exemplo da abstração da representação do mapeamento pode ser observada na Figura 5.16. Onde, a primeira coluna refere-se ao nome da lista, a segunda coluna ao *URI* da ontologia de domínio, e a terceira coluna indica qual é a classe correspondente na ontologia.

Nome.lst	*URI*	Classe

Figura 5.16: Abstração do mapeamento das listas para a classe correspondente na ontologia.

Um exemplo, conforme a primeira linha do arquivo *mapping.def*, pode ser visualizado na Figura 5.17.

Nesse exemplo, a lista é *"dialeto.lst"*, o *URI* da ontologia de domínio é *"http://www.owl-ontologies.com/InstrumentoLinguistico.owl"*, e a classe na ontologia é *"Dialeto"*.

dialeto.lst	http://www.owl-ontologies.com/InstrumentoLinguistico.owl	Dialeto

Figura 5.17: Mapeamento da lista: dialeto.lst, URI: http://www.owl-ontologies.com/InstLinguistico.owl e Classe: Dialeto.

5.3.2 Construção das Regras

Esse processo é realizado pelo Engenheiro do Sistema com colaboração do Especialista do Domínio para a construção de regras que permitam atribuir nomes semânticos às marcações que serão realizadas no documento anotado. Essas regras devem permitir que sejam marcados adequadamente nos documentos os conceitos primitivos e os conceitos derivados solicitados pelo especialista.

A etapa de construção das regras, conforme Figura 5.5, é alimentada pelo Conjunto de Requisitos de Conceitos Derivados, pelas Listas de Conceitos Primitivos e pela Lista de Mapeamento dos Conceitos, produzidos na etapa de Construção do Dicionário. Essa etapa tem como controle a ontologia de domínio e o recurso JAPE que determina a gramática da regra. Para auxiliar a tarefa de construção das regras, podem ser realizadas consultas SPARQL sobre a ontologia de domínio e essas consultas podem ser executadas na ferramenta Protégé no painel SPARQL *Query*.

Foram estabelecidos dois conjuntos com quatro tipos de regra que atendem o objetivo de anotar semanticamente os documentos de acordo com o estipulado pelo especialista de domínio no objeto de estudo apresentado nesta dissertação. Esses tipos são:

- Regras para Conceitos Primitivos:
  - Conceitos Primitivos mapeados para a Ontologia de Domínio.
- Regras para Conceitos Derivados:
  - Relacionamento SuperClasse e Subclasse.
  - Relacionamento Classe e Propriedade.
  - Relacionamento Classes e Conceitos Derivados.

Conceitos primitivos mapeados para a Ontologia de Domínio

Primeiramente, devem ser estabelecidas regras que permitam identificar adequadamente cada anotação de conceito primitivo para uma classe correspondente na ontologia. Para isso, essa regra deve ser baseada no mapeamento das listas para o conceito-ontologia, conforme abstração apresentada na Figura 5.16.

Conforme a discussão do mapeamento das listas para o conceito-ontologia, realizada na seção 5.3.1, a última coluna representa a classe na ontologia mais próxima do conceito a ser anotado. A informação presente nessa coluna é utilizada na regra para indicar qual classe é mencionada na anotação.

Como pode ser observado na Figura 5.18, a regra exemplificada é uma abstração de como devem ser as regras de mapeamento de cada menção do conceito primitivo para a classe correspondente na ontologia. Essa é a fase de Mencao representada na linha 4. A entrada da regra são as marcações nos documentos referentes ao dicionário *(Lookup)*, linha 5. E na linha 7, a regra é denominada de **MencaoClasse**.

As informações contidas no arquivo *"mapping.def"* atribuem uma propriedade indicando qual classe corresponde a cada marcação de conceito primitivo. Assim, na linha 9, essa regra busca nas marcações do tipo *Lookup* aquelas que atendam ao padrão estabelecido para a propriedade *class*, isto é, que sejam iguais a **Classe**. E as atribui à variável **classe** na linha 10. Esse é o lado esquerdo da regra, como denominado no JAPE (LHS). Nesse lado é onde são estabelecidos os padrões a serem buscados nos documentos. Depois da seta, ou ao lado direito (RHS), todas as ocorrências encontradas e atribuídas à variável **classe** recebem uma marcação do tipo *Mention* que faz uma menção da **Classe** na ontologia de domínio, linha 13.

mencao.jape

```
1 // Regra abstrata que associa um conceito primitivo
2 // a uma classe na ontologia de domínio.
3
4   Phase: Mencao
5   Input: Lookup
6
7   Rule: MencaoClasse
8   (
9     {Lookup.class == Classe}
10  ):classe
11
12  -->
13   :classe.Mention = {class = :classe.Lookup.class}
```

Figura 5.18: Arquivo JAPE com a regra abstrata que faz o mapeamento do conceito primitivo para a classe na ontologia de domínio.

Essa regra deve ser construída para cada conceito primitivo existente no mapeamento representado pelo arquivo *"mapping.def"*, que pode ser visualizado na Figura 5.15.

Por exemplo, a regra para o mapeamento da primeira linha do arquivo *"mapping.def"* pode ser visualizada na Figura 5.19. A fase e a entrada serão iguais para todas as regras desse tipo, *Mencao* e *Lookup*, respectivamente nas linhas 4 e 5.

Na linha 7, tem-se o nome da regra **MencaoDialeto**. Aqui foi utilizado o nome padrão estabelecido pela regra abstrata acrescido do nome da classe em questão, nesse caso, Dialeto.

Na linha 9, o padrão buscará por ocorrências de classe na propriedade *Lookup.class* que sejam iguais a classe Dialeto. As ocorrências encontradas são atribuídas à variável **dialeto** na linha 10. Na linha 13, todas as ocorrências que foram atribuídas a essa variável recebem uma marcação do tipo *Mention*, que faz uma menção da classe **Dialeto** na Ontologia InstrumentoLinguistico.

mencaoDialeto.jape

```
1 // Regra MencaoDialeto que associa um conceito primitivo dialeto
2 // a classe Dialeto na ontologia de domínio.
3
4   Phase: Mencao
5   Input: Lookup
6
7   Rule: MencaoDialeto
8   (
9     {Lookup.class == Dialeto}
10  ):dialeto
11
12  -->
13   :dialeto.Mention = {class = :dialeto.Lookup.class}
```

Figura 5.19: Arquivo JAPE com a regra que faz o mapeamento de uma ocorrência de termos relacionados ao conceito dialeto para a classe Dialeto na Ontologia InstrumentoLinguistico.

Outra maneira de anotar semanticamente os conceitos primitivos, é atribuir a cada menção uma anotação com um tipo que corresponde ao nome da classe que pertence, para isso, são construídas regras que seguem o modelo abstrato representado na Figura 5.20.

Classe.jape

```
1 // Regra abstrata Classe que associa um conceito primitivo
2 // a uma classe NomedaClasse na ontologia de domínio.
3
4   Phase: Classe
5   Input: Mention
6
7   Rule: ClasseNomedaClasse
8   (
9     {Mention.class == NomedaClasse}
10  ):nomeclasse
11
12  -->
13   :nomeclasse.NomedaClasse = {class = :nomeclasse.Mention.class}

C ▾  Largura das tabulações: 8 ▾  Lin 2, Col 9  INS
```

Figura 5.20: Arquivo JAPE com a regra abstrata que faz uma anotação especializada do conceito primitivo para a classe na ontologia de domínio.

Na Figura 5.21, é possível observar a regra para anotação de classe do tipo *Lingua*. Na linha 7, tem-se o nome da regra **ClasseLingua**. Aqui foi utilizado o nome padrão estabelecido pela regra abstrata acrescido do nome da classe em questão, nesse caso, Lingua. Na linha 9, o padrão buscará por ocorrências de classe na propriedade *Mention.class* que sejam iguais a classe Lingua. As ocorrências encontradas são atribuídas à variável **lingua** na linha 10. Na linha 13, todas as ocorrências que foram atribuídas a essa variável recebem uma marcação do tipo *Lingua*, que faz uma menção da classe **Lingua** na Ontologia InstrumentoLinguistico.

Esse tipo de anotação pode ser realizada para todas as classes, ou apenas para as que contribuem na anotação de conceitos derivados mais complexos.

ClasseLingua.jape

```
1 // Regra ClasseLingua que associa um conceito primitivo lingua
2 // a classe Lingua na ontologia de domínio.
3
4   Phase: Classe
5   Input: Mention
6
7   Rule: ClasseLingua
8   (
9     {Mention.class == Lingua}
10  ):lingua
11
12  -->
13   :lingua.Lingua = {class = :lingua.Mention.class}
```

C ▾ Largura das tabulações: 8 ▾ Lin 4, Col 16 INS

Figura 5.21: Arquivo JAPE com a regra ClasseLingua que faz a anotação do tipo *"Lingua"* dos conceitos primitivos para a classe "Lingua" na ontologia de domínio.

Após a construção das regras para os conceitos primitivos, o engenheiro do sistema junto ao especialista de domínio, realizam uma prospecção na(s) ontologia(s) buscando os relacionamentos entre esses conceitos primitivos de forma a tentar chegar nos conceitos derivados solicitados pelo especialista de domínio, conforme consta no Conjunto de Requisitos de Conceitos Derivados. O engenheiro do sistema, com apoio da ferramenta Protégé e da linguagem SPARQL, constrói consultas para investigar os axiomas e os relacionamentos existentes na ontologia e que atendam aos requisitos estabelecidos pelo especialista.

Para anotar os conceitos derivados devem ser criadas regras junto ao especialista de domínio, e que atendam aos requisitos refinados de conceito derivado. Os conceitos derivados podem ser extraídos da ontologia através de consultas SPARQL, e os relacionamentos encontrados influenciam na construção das regras para anotação. As regras desenvolvidas para realizar a anotação dos conceitos derivados nos documentos são apresentadas nos próximos parágrafos.

Relacionamento SuperClasse e Subclasse

Essas são as regras que associam as subclasses à sua superclasse de acordo com a ontologia de domínio.

Na Figura 5.22, pode ser observada a regra **SuperClasse** que é uma abstração de como podem ser feitas as regras que associam uma ou mais subclasses à sua superclasse corresponde conforme a ontologia de domínio. A entrada da regra são as marcações nos documentos referentes ao mapeamento do conceito primitivo para a classe corresponde na ontologia, realizadas pela regra descrita anteriormente e identificadas pela anotação *(Mention)*, linha 4.

superclasse.jape

```
1 // Regra abstrata que associa as subclasses à sua superclasse
2
3  Phase:superclasse
4  Input: Mention
5
6   Rule: SuperClasse
7   (
8     {Mention.class == SubClasse1} | {Mention.class == SubClasse2} | ... | {Mention.class == SubClasseN}
9   ):superclasse
10
11  -->
12   :superclasse.NomeSuperClasse = {superclasse = NomeSuperClasse, subclasse = :superclasse.Mention.class}
```

Figura 5.22: Arquivo JAPE com a regra abstrata que associa a(s) subclasse(s) à sua superclasse.

Ainda de acordo com a Figura 5.22, na linha 6, tem-se o nome da regra que é **SuperClasse**, e o que se busca nessa regra são as ocorrências das classes que correspondem ao padrão descrito na linha 8. Isto é, quando a marcação *(Mention.class)* for igual a SubClasse1 **ou**, representado por uma barra (|), for igual a SubClasse2 **ou**, for igual a SubClasseN . Essas ocorrências encontradas são atribuídas à variável **superclasse**, linha 9. Na linha 12, as ocorrência atribuídas a essa variável recebem a marcação do tipo *NomeSuperClasse*, e o **NomeSuperClasse** é atribuído para a propriedade **superclasse**, além disso, o devido nome de classe de cada ocorrência é atribuído para a propriedade **subclasse**.

Um exemplo para essa regra, conforme o domínio abordado nesta dissertação, pode ser visualizado na Figura 5.23.

superclasseAssunto.jape

```
1 // Regra que associa as subclasses Dialeto, Idioma e Lingua à superclasse Assunto
2
3  Phase:superclasse
4  Input: Mention
5
6   Rule: SuperClasseAssunto
7   (
8     {Mention.class == Dialeto} | {Mention.class == Idioma} | {Mention.class == Lingua}
9   ):superclasse
10
11  -->
12   :superclasse.Assunto = {superclasse = Assunto, subclasse = :superclasse.Mention.class}
```

Figura 5.23: Arquivo JAPE com a regra que associa as subclasses Dialeto, Idioma e Lingua à superclasse Assunto.

A fase e a entrada serão iguais para todas as regra desse tipo, *superclasse* e *Mention*, respectivamente nas linhas 3 e 4.

Na linha 6, tem-se o nome da regra **SuperClasseAssunto**, aqui foi utilizado o nome padrão estabelecido pela regra abstrata acrescido do nome da superclasse em questão, nesse caso, Assunto. Na linha 8, o padrão buscará por ocorrências de classe na marcação *Mention.class* que sejam iguais a Dialeto, ou a Idioma, ou a Lingua. As ocorrências encontradas são atribuídas

à variável **superclasse**. Na linha 12, todas as ocorrências que foram atribuídas a essa variável recebem uma marcação do tipo *Assunto*, a propriedade **superclasse** recebe o nome **Assunto**, e o nome de classe correspondente a cada ocorrência é atribuído para a propriedade **subclasse**.

Relacionamento Classe e Propriedade

Nesse tipo de regra a ideia é relacionar uma classe à alguma propriedade relacionada. É importante considerar que os resultados buscados com essa regra podem ser relevantes para formar conceitos derivados mais complexos.

Por exemplo, para especificar a classe Povos e selecionar apenas os conceitos que indicam que o povo ou naturalidade é brazileiro (com suas distintas grafias), pode ser utilizada a regra representada na Figura 5.24.

Naturalidade_brazileiro.jape

```
1 //Regra para o conceito tem_Naturalidade brazileiro
2
3  Phase: Naturalidadebrazileiro
4  Input: Mention Token
5
6   rule: Naturalidade_brazileiro
7   (
8     {Mention.class == Povos,
9       Token.string=~"bra"}
10  ):povo
11
12  -->
13   :povo.Naturalidade_brazileiro = {Natural_de = Brazil, rule = "Naturalidade_brazileiro" }
```

Figura 5.24: Arquivo JAPE com a Regra que especifica a classe Povos em Naturalidade_brazileiro.

Nessa regra, a fase é denominada de Naturalidadebrazileiro na linha 3, e esse nome indica qual a relação que a regra vai marcar nos documentos. Na linha 4, a regra recebe como entrada as marcações *Mention* e *Token*. O nome da regra **Naturalidade_brazileiro** , na linha 6, é praticamente o mesmo da fase pois indica o objetivo de marcação da regra. Nas linhas 8 e 9, a regra vai buscar nas menções da classe Povos ocorrências que iniciam com a string **"bra"**, os resultados encontrados são atribuídos para a variável **povo**, na linha 10. Na linha 13, cada ocorrência atribuída à variável **povo** recebe a marcação *Naturalidade_brazileiro* e a propriedade de anotação *Natural_de = Brazil*, essa propriedade foi extraída da ontologia de domínio. Por fim, a propriedade *rule = "Naturalidade_brazileiro"* que faz referência à regra que realizou a marcação.

Relacionamento Classes e Conceitos Derivados

Como foi discutido no capítulo 4, as relações entre os conceitos primitivos presentes na ontologia formam os conceitos derivados. Por exemplo, para expressar o conceito derivado

"Dialecto brazileiro" de acordo com a Ontologia InstrumentoLinguistico tem-se, Dialeto é uma classe relacionada com a instância brazileiro da classe Povos através do relacionamento tem_Naturalidade (conforme explicado na seção 4.2):

Dialeto *tem_Naturalidade* **brazileiro**

A regra de anotação desse conceito derivado pode ser observado na Figura 5.25.

dialetotemNaturalidadebrazileiro.jape

```
1 //Regra para o conceito derivado "dialeto brazileiro"
2
3  Phase: DialetoBrazileiro
4  Input: Mention Naturalidade_brazileiro
5
6   rule: Dialeto_brazileiro
7   (
8    ({Mention.class==Dialeto}):superconcept
9    ({Naturalidade_brazileiro}):subconcept
10  ):dialeto
11
12  -->
13   :dialeto.Dialeto_tem_Naturalidade_brazileiro = { rule = "Dialeto_brazileiro" },
14   :superconcept.Domain = { rule = "Dialeto_brazileiro" },
15   :subconcept.Range = { rule = "Dialeto_brazileiro" }
```

Figura 5.25: Arquivo JAPE com a Regra que identifica o conceito derivado "dialeto brazileiro".

Essa regra, que anota as ocorrências do conceito derivado "Dialeto brazileiro", tem como entrada, na linha 4, as marcações do tipo *Mention* e Naturalidade_brazileiro, que é outro conceito derivado já especificado pela regra anteriormente apresentada. O nome da regra é **Dialeto_brazileiro**, linha 6. Na linha 8, a regra busca entre as menções de classe quais correspondem a classe "Dialeto", e as atribui à variável **superconcept**. Em seguida, na linha 9, a regra busca por ocorrências de *Naturalidade_brazileiro* no *token* subsequente ao primeiro padrão estabelecido, e as atribui à variável **subconcept**. As ocorrências encontradas que atendam aos dois padrões são atribuídas à variável **dialeto** na linha 10.

Ao lado direito da regra, na linha 13, as ocorrências da variável **dialeto** recebem uma marcação do tipo *Dialeto_tem_Naturalidade_brazileiro*. Na linha 14, os valores atribuídos a variável **superconcept** recebem a marcação do tipo *Domain*. Na linha 15, os valores atribuídos a variável **subconcept** recebem a marcação do tipo *Range*. Além disso, as marcações das linhas 13, 14 e 15, recebem uma propriedade que indica qual regra realizou a anotação, nesse caso, *rule = "Dialeto_brazileiro"*.

Outro exemplo de regra para anotação semântica em conceitos derivados ainda mais complexos, pode ser observado na Figura 5.26.

Esse tipo de regra é estabelecida junto com o especialista de domínio, pois envolve a busca por mais de um conceito derivado. Os conceitos derivados buscados apresentam características similares e podem ser encontrados por um padrão mais expressivo e determinante na identificação do domínio em questão.

Para auxiliar na descoberta dessas possíveis regras para anotação mais complexa, é recomendável que o engenheiro do sistema aplique os tipos de regras explicados anteriormente em uma amostra do córpus para prover ao especialista de domínio os documentos anotados semanticamente. Assim, pode permitir a observação e possível descoberta de novas regras mais complexas que o especialista de domínio deseje montar em função dos conceitos derivados já expressos em regras.

Por exemplo, a regra apresentada na Figura 5.26.

AssuntoModalidadeLusofonos.jape

```
1 //Regra que encontra ocorrências de Assunto seguido por Modalidade que pode ser Falada ou Escrita,
2 // e depois por algum País Lusófono.
3
4  Phase: AssuntoModalidadePaisesLusofonos
5  Input: Mention Token Modalidade Naturalidade_lusofono
6
7   rule: Assunto_Modalidade_Paises_Lusofonos
8   (
9     ({Assunto}|{Naturalidade_lusofono})* :superconcept
10    ({Modalidade})
11    ({Token.string=="no"} | {Token.string=="em"} | {Token.string=~"d"})*
12    ({Mention.class==Paises_Lusofonos}):subconcept
13  ):assuntomodolusofono
14
15  -->
16   :assuntomodolusofono.Assunto_Modalidade_Paises_Lusofonos = { rule = "Assunto_Modalidade_Paises_Lusofonos" },
17   :superconcept.Domain = { rule = "Assunto_Modalidade_Paises_Lusofonos" },
18   :subconcept.Range = { rule = "Assunto_Modalidade_Paises_Lusofonos" }
```

Figura 5.26: Arquivo JAPE com a Regra que identifica as ocorrências de vários conceitos derivados correlacionados.

De acordo com o especialista, e após investigação nos documentos que compõem a amostra anotada do córpus, encontrou-se um padrão recorrente e abrangente que determina que um Assunto é seguido de uma Modalidade e seguido de um País Lusófono. Esse padrão é encontrado em vários conceitos derivados correlacionados, e garante expressividade para determinar o domínio da constituição da Língua Portuguesa no Brasil. Com esse padrão busca-se encontrar conceitos derivados, tais como: “língua falada no Brazil”, “portuguez falado no Brazil”, “portuguez falado em Portugal”, “falar de Portugal”, entre outros.

A fase da regra é AssuntoModalidadePaisesLusofonos, linha 4, e as entradas são os tipos *Mention*, *Token*, Modalidade e Naturalidade_lusofono, representados na linha 5. O nome da regra é Assunto_Modalidade_Paises_Lusofonos, conforme a linha 7.

Da linha 9 até a 12, está representado o padrão estabelecido. Na linha 9, a regra busca por ocorrências marcadas com a classe Assunto **ou**, representado por uma barra (|), com

o conceito derivado Naturalidade_lusofono. Porém, nessa linha depois do fechamento dos parênteses, existe um símbolo asterisco (∗) que é o operador que permite marcar zero ou mais vezes o padrão descrito, pois a classe Assunto pode ser oculta. As ocorrências encontradas são atribuídas a variável **superconcept**.

Na linha 10, a regra busca no *token* seguinte, por uma ocorrência da classe Modalidade que pode ser falada ou escrita, com as devidas variantes de grafia.

Na linha 11, a regra busca no próximo *token* por alguma ocorrência com marcação do tipo *Token* cuja propriedade *string* seja igual a "no", ou (|) "em", ou (|) que inicie com a letra "d". Aqui ocorre novamente a presença do operador (∗) que permite a marcação de zero ou mais ocorrências desse padrão.

Na linha 12, a regra busca no próximo *token* por alguma menção de classe correspondente a classe "Paises_Lusofonos", e a atribui à variável **subconcept**. Por fim, na linha 13, todas as ocorrências de conceito derivado que atendam ao padrão estabelecido são atribuídas à variável **assuntomodolusofono**.

Após a seta na linha 15, ao lado direito da regra, tem-se as instruções de qual o tipo de anotação é atribuído para cada variável. Na linha 16, as ocorrências da variável **assuntomodolusofono** recebem a marcação do tipo *Assunto_Modalidade_Paises_Lusofonos*. Na linha 17, as ocorrências da variável **superconcept** recebem a marcação do tipo *Domain*. E na linha 18, as ocorrências da variável **subconcept** recebem a marcação do tipo *Range*. Além disso, as marcações da linha 16, 17 e 18, recebem uma propriedade que indica qual regra realizou a anotação, nesse caso, *rule = "Assunto_Modalidade_Paises_Lusofonos"*.

Lembrando que, para determinar as regras é de extrema importância o trabalho desenvolvido em parceria com o especialista de domínio. Especialmente nesse caso, em que o intuito é de construir regras que buscam diversos conceitos derivados, pois somente o especialista pode dar o aval de quais relações são as mais expressivas e como elas se combinam para poder formar um padrão que representa esses conceitos derivados.

Essa tarefa é custosa e não trivial, pois exige o trabalho constante de refinamento da regra por parte do engenheiro do sistema e do especialista de domínio. Porém, quando estabelecida, gera resultados expressivos para a recuperação de informação, já que, a partir de uma única regra são localizadas e marcadas inúmeras ocorrências de diversos conceitos derivados investigados.

5.4 Extração de Informação Baseada em Ontologia

A arquitetura do processo automático de extração de informação baseada em ontologia (OBIE), está representada na Figura 5.27. O processo de OBIE é dividido em duas etapas: a Localização dos Conceitos e a Anotação Semântica.

O processo tem como entrada um córpus composto por documentos no formato XML tokenizado, ou XML POSTagger, caso tenha sido implementada a etapa de análise morfossintática. Além dos artefatos construídos no processo anterior, que são o Dicionário, o Mapeamento e as Regras, conforme apresentado na Figura 5.1.

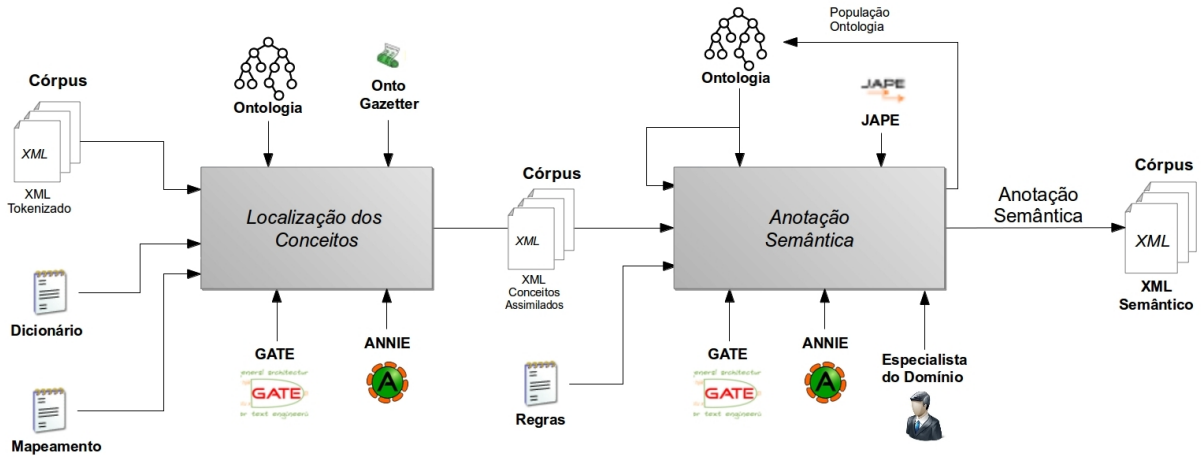


Figura 5.27: Arquitetura da Extração de Informação Baseada em Ontologia.

O córpus passa pela etapa de Localização de Conceitos, em que, as tarefas de extração de informação baseadas em dicionário são aplicadas ao córpus para localização das ocorrências de conceitos primitivos da ontologia de domínio. Nessa etapa, também é realizada a vinculação dos conceitos localizados no córpus à classe correspondente na ontologia usando o Mapeamento. O córpus contendo os documentos no formato XML Conceitos Assimilados, é a saída dessa etapa e entrada da etapa de Anotação Semântica.

A etapa de Anotação Semântica é o processo onde são reconhecidos e anotados os conceitos primitivos e os conceitos derivados, por meio de regras baseadas na ontologia e, desenvolvidas pelo Engenheiro do Sistema com supervisão do Especialista de Domínio.

O resultado dessa etapa, é um córpus formado por documentos no formato XML, o XML semântico, que possui as anotações dos conceitos primitivos e conceitos derivados extraídos da

ontologia de domínio. Além disso, é possível popular a ontologia com as instâncias identificadas como uma menção de classe em cada documento do córpus. Nas próximas seções cada etapa será detalhada.

5.4.1 Localização dos Conceitos

A localização das ocorrências dos conceitos primitivos encontrados nos documentos do córpus é realizada com o auxílio de um dicionário formado por um conjunto de listas que contêm os conceitos primitivos presentes na ontologia do domínio.

A entrada dessa etapa são os documentos no formato XML tokenizado, ou XML POSTagger, as listas que representam o dicionário e o mapeamento, construídos na etapa de Construção dos Artefatos. O OntoGazzeter é o recurso de controle que junto ao Gate, realiza o processo de localizar os conceitos primitivos.

Nessa etapa, o recurso OntoGazzeter utiliza o dicionário representado pelo arquivo *"lists.def"* para realizar as anotações do tipo *Lookup* nas ocorrências encontradas nos documentos de acordo as listas. Além disso, cada anotação especifica suas propriedades de maior tipo, de menor tipo e a propriedade tipo *class* que indica qual a classe correspondente na ontologia de domínio. As especificações para as propriedades maior tipo e menor tipo estão no arquivo *"lists.def"*, e para a propriedade tipo *class* no arquivo *"mapping.def"*.

Um exemplo da anotação do tipo *Lookup* realizada pela aplicação AnotacaoSemantica pode ser visualizada na interface do GATE, Figura 5.28.

Ao lado esquerdo da figura, é possível visualizar os componentes utilizados no GATE. A aplicação AnotacaoSemantica, os recursos de linguagem "O Dialecto Brazileiro" que é o documento utilizado como exemplo, o córpus CorpusRevistaBrazileira que contém uma amostra dos documentos pertencentes a Revista Brazileira, e a Ontologia InstrumentoLinguistico. Em seguida, os recursos de processamento que implementam a abordagem, como o recurso Dicionário_OntoGazzeter utilizado para realizar a anotação do tipo *Lookup*.

Na parte central da figura, pode ser observado um trecho do documento "O Dialecto Brazileiro" com as anotações do tipo *Lookup*, grifadas com a cor magenta, nas ocorrências correspondentes ao conteúdo presente nas listas que formam o dicionário. A anotação *Lookup* pode ser selecionada ao lado direito da figura quando a aba *Annotation Sets* é ativada, e isso permite a visualização das ocorrências localizadas.

Por fim, na parte inferior são listadas as informações de cada anotação caso a aba *Annotation*

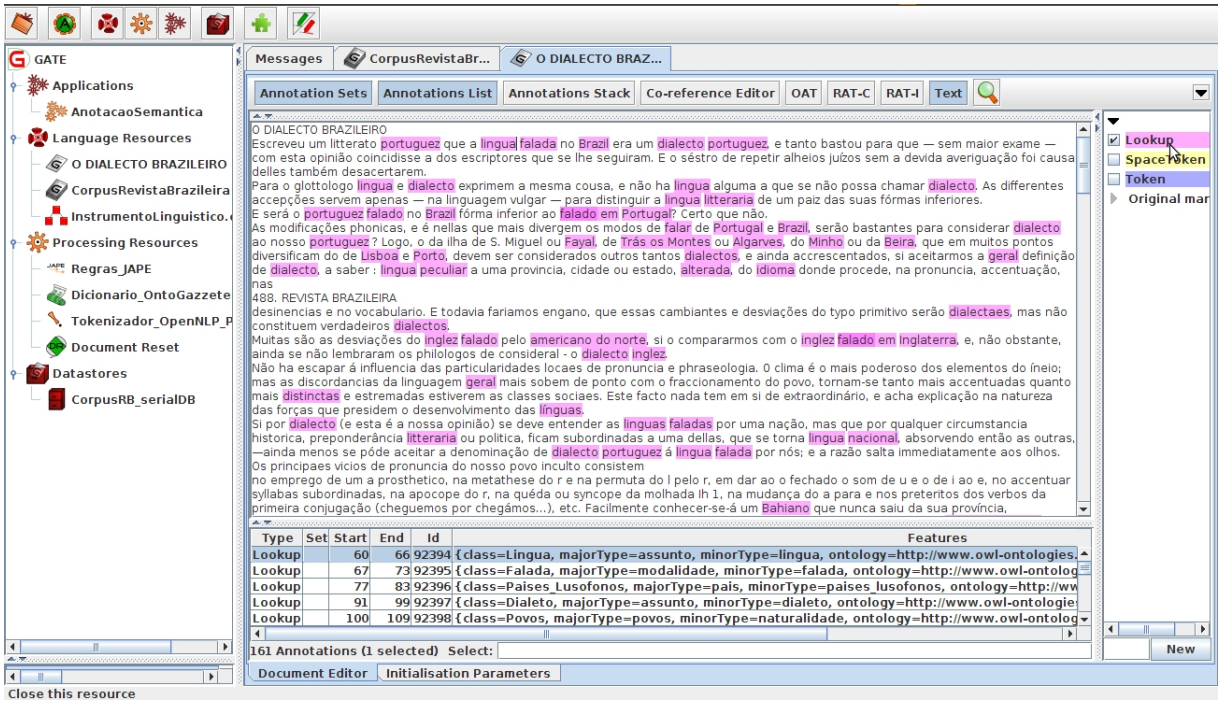


Figura 5.28: Anotações do tipo *Lookup* realizadas pela aplicação AnotacaoSemantica no GATE.

List esteja ativada. As informações identificam o tipo, o *id* início e o *id* fim da ocorrência, o *Id* da anotação e as características de anotação, que neste caso são *class*, *majorType*, *minorType* e *ontology*. Logo em seguida, é possível visualizar o número total de anotações realizadas para o tipo selecionado.

A anotação grifada listada na primeira linha, com *id* início = 60 e o *id* fim = 66, e as anotações das linhas 4 e 5, com *id* início = 91 e o *id* fim = 99 e com *id* início = 100 e o *id* fim = 109, são utilizadas nos exemplos explicados a seguir.

Na Figura 5.29, é possível visualizar os detalhes de uma anotação do tipo *Lookup* em uma ocorrência de "lingua" no documento "O Dialecto Brazileiro".

O conceito primitivo "lingua" é representado pela lista *lingua.lst* e de acordo com o arquivo *"lists.def"* a lista tem como valores para a propriedade maior tipo *(majorType)*: "assunto"; e para a propriedade menor tipo *(minorType)*: "lingua". Os valores podem ser observados na janela *pop-up* que representa a anotação. Além disso, através do arquivo *"mapping.def"*, a lista *lingua.lst* tem como valor para a propriedade *(class)* a classe "Lingua". A propriedade *(ontology)* tem como valor a URI correspondente da Ontologia InstrumentoLinguistico.

Essa anotação é realizada em cada ocorrência encontrada nos documentos que remete ao conteúdo presente nas listas. Em cada anotação, são atribuídos os valores para as propriedades considerando o que consta nos arquivos *"lists.def"* e *"mapping.def"*, conforme apresentado no

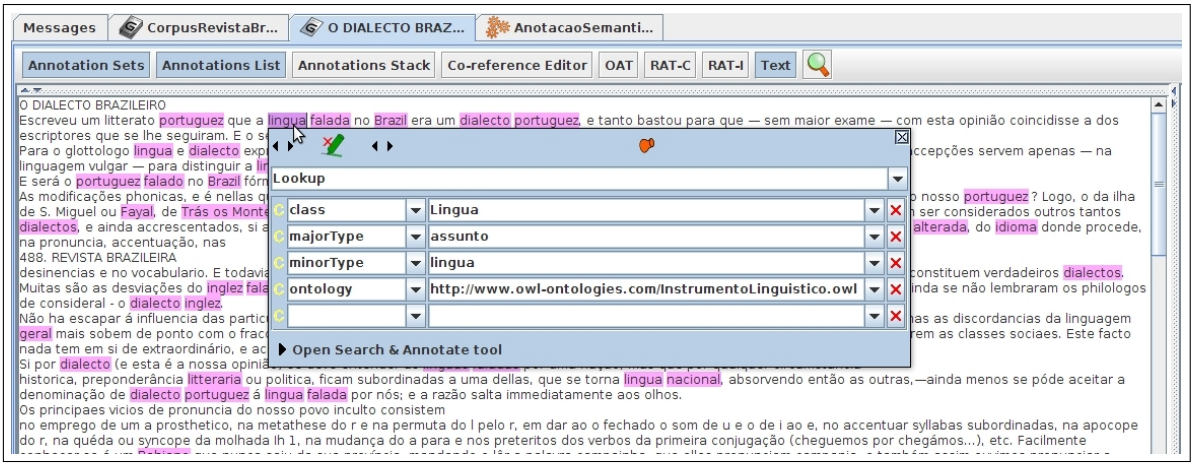


Figura 5.29: Detalhes da anotação do tipo *Lookup* em uma ocorrência de "lingua" no documento "O Dialecto Brazileiro".

exemplo anterior.

O resultado dessa etapa é um córpus com documentos no formato XML Conceitos Assimilados. Um exemplo do documento XML pode ser observado na Figura 5.30. Nessa figura, pode ser visualizada a representação do documento XML com as anotações do tipo *Lookup* em um trecho do documento "O Dialecto Brazileiro". As anotações estão representadas pela marcação *<Lookup>* **ocorrência** *</Lookup>*, em cada ocorrência localizada.

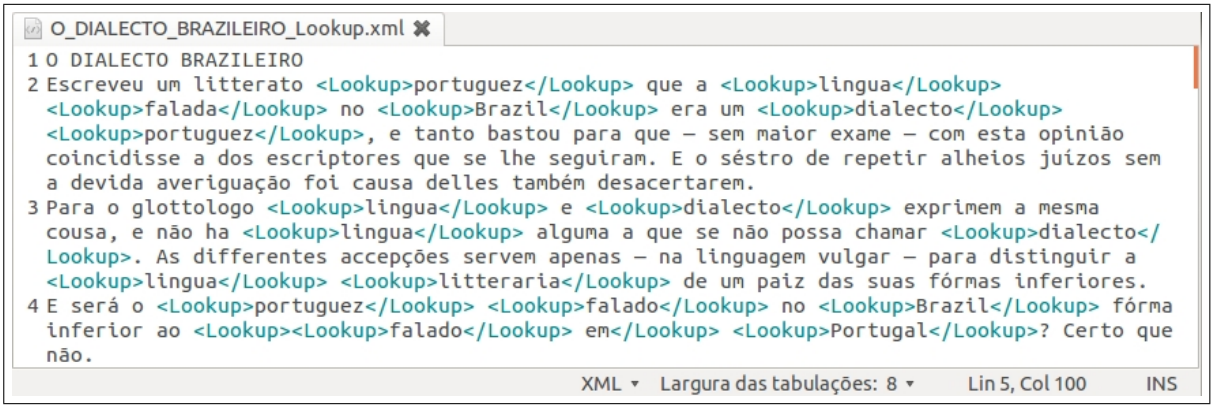


Figura 5.30: Documento XML com anotações do tipo *Lookup* em um trecho do documento "O Dialecto Brazileiro".

O exemplo de um documento XML com a formatação definida pelo GATE pode ser observado na Figura 5.31. Nessa figura os *tokens* são marcados por nodos serializados, isto é, cada *token* tem *ids* que indicam o seu início e o seu fim. Para exemplificar como são feitas as anotações no documento XML formatado, pode ser observado o trecho grifado que representa o conceito derivado "dialecto portuguez", no início do documento "O Dialecto Brazileiro".

A anotação identifica o *token* "dialecto" inicialmente pelo *<Node id = "91">* e no fim pelo

<*Node id* = *"99"*>. E o *token* "portuguez" inicialmente pelo <*Node id* = *"100"*> e no fim pelo <*Node id* = *"109"*>.

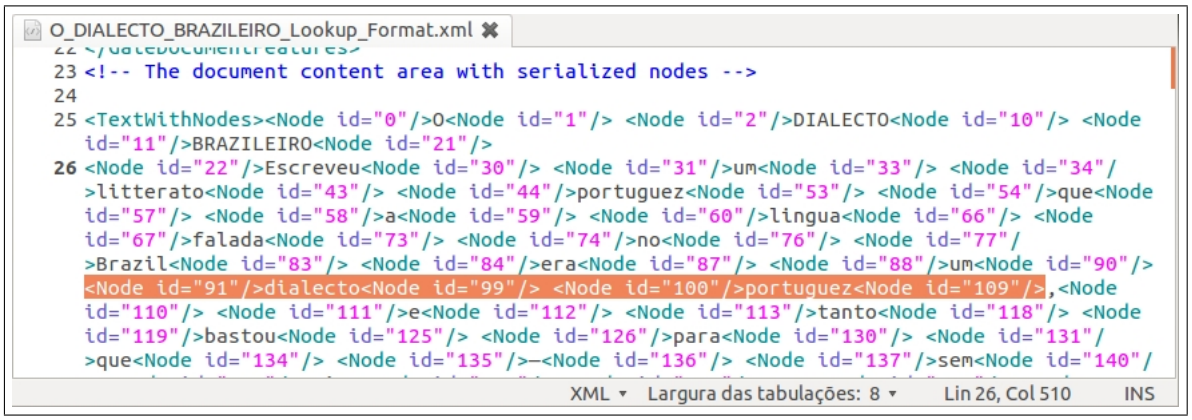
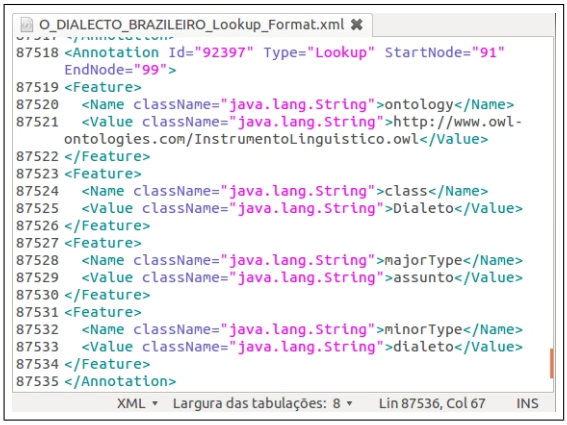


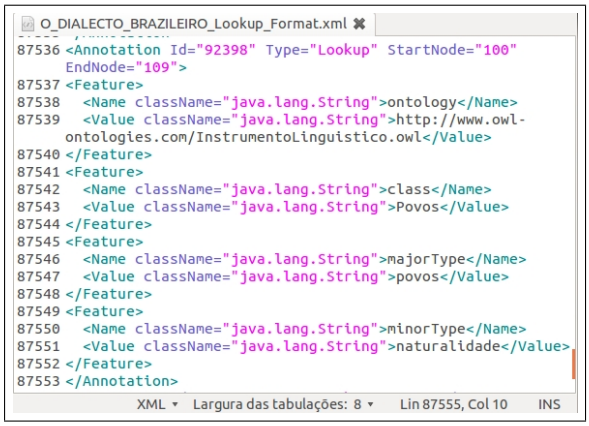
Figura 5.31: Documento XML com anotação em forma de nodos seriados de um trecho do documento "O Dialecto Brazileiro".

No final do documento XML formatado, as características de cada anotação são atribuídas ao *id* correspondente. Por exemplo, na Figura 5.32 (a) podem ser observadas as características da anotação *Lookup* com *Id* = "92397", a partir da linha 87518, para o *token* cujo nodo tenha início em 91 e término em 99, isto é, para o *token* "dialecto". E na Figura 5.32 (b), as características da anotação *Lookup* com *Id* = "92398", a partir da linha 87536, para o *token* cujo nodo tenha início em 100 e término em 109, isto é, para o *token* "portuguez".

Essas características são atribuídas para todas as ocorrências de anotação do tipo *Lookup*, conforme especificações estipuladas anteriormente.



(a) Anotação *Id* = "92397"



(b) Anotação *Id* = "92398"

Figura 5.32: Documento XML com as características de anotação dos nodos que correspondem ao conceito derivado "dialeto portuguez" no início do documento "O Dialecto Brazileiro".

O córpus contendo os documentos no formato XML Conceitos Assimilados é a entrada para a próxima etapa onde são feitas as anotações semânticas.

5.4.2 Anotação Semântica

A Anotação Semântica é a etapa de anotação das ocorrências dos conceitos primitivos e conceitos derivados encontrados nos documentos do córpus, de acordo com a ontologia de domínio. Nessa etapa, são utilizadas as regras construídas no processo de Construção dos Artefatos, a fim de localizar os conceitos primitivos e conceitos derivados nos documentos e anotá-los adequadamente.

Conforme exposto na Figura 5.1, a entrada dessa etapa são os documentos no formato XML Conceitos Assimilados, as Regras e a ontologia de domínio. O recurso de processamento JAPE e a ontologia de domínio são os controles do processo que junto ao mecanismo Gate, permitem a anotação semântica dos conceitos primitivos e dos conceitos derivados. O processo é automático, no entanto, o Especialista de Domínio pode interferir na fase de População da Ontologia atribuindo corretamente as instâncias identificadas pelo processo. O resultado dessa etapa é um córpus com documentos no formato XML Semântico.

Nos próximos parágrafos são apresentados alguns exemplos da execução da etapa de Anotação Semântica dos Conceitos Primitivos e dos Conceitos Derivados.

Anotação Semântica dos Conceitos Primitivos

A anotação dos conceitos primitivos é realizada para vincular à ontologia de domínio as ocorrências localizadas pela marcação do tipo *Lookup*. Para isso, são construídas regras conforme o modelo de regra abstrata **MentionClasse**, representado na Figura 5.18. Essas regras atribuem uma anotação do tipo *Mention* para cada uma das ocorrências localizadas.

Para ilustrar como isso acontece na interface do Gate, pode ser observada a Figura 5.33.

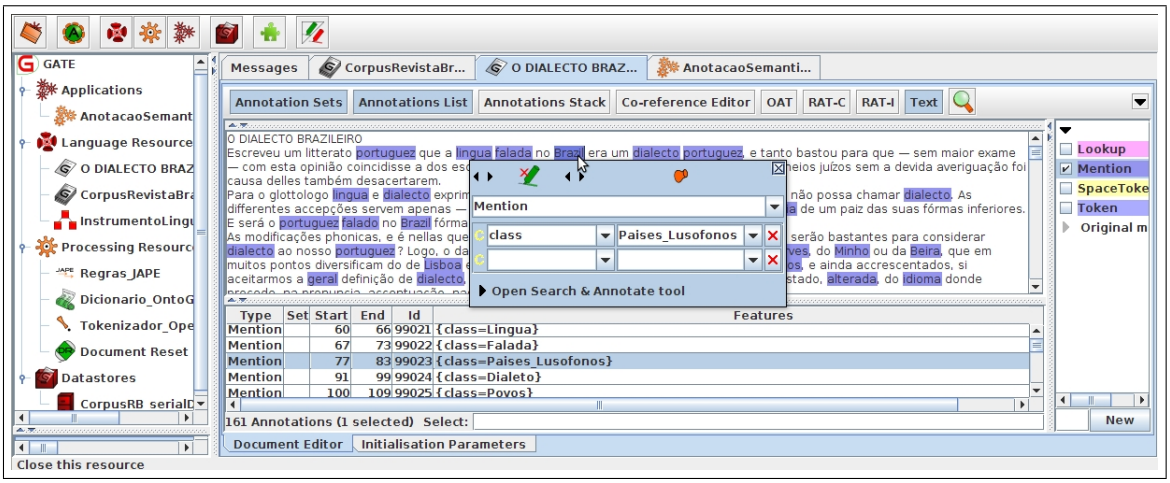


Figura 5.33: Detalhes da anotação do tipo *Mention* em uma ocorrência de "Brazil" no documento "O Dialecto Brazileiro".

Nessa figura, é possível visualizar as anotação do tipo *Mention* no mesmo trecho do documento "O Dialecto Brazileiro" utilizado anteriormente. Por exemplo, na primeira linha do documento o quarto token anotado é uma ocorrência de "Brazil". De acordo com a Ontologia InstrumentoLinguistico, "Brazil" é uma instância da classe Paises_Lusofonos, por isso, recebe como anotação o tipo *Mention* e a propriedade *class* Paises_Lusofonos. Esse tipo de anotação é realizada em todas as ocorrências localizadas respeitando a informação determinada pela propriedade *class* da anotação *Lookup*.

Na figura Figura 5.34, é possível observar algumas ocorrências de anotações do tipo *Mention* vinculadas à Ontologia InstrumentoLinguistico, que pode ser visualizada parcialmente ao lado direito da figura pois a *tab* OAT foi ativada. O cursor aponta para a classe Paises_Lusofonos que corresponde a classe da ocorrência "Brazil". Todas as marcações na cor vermelha correspondem a classe Paises_Lusofonos.

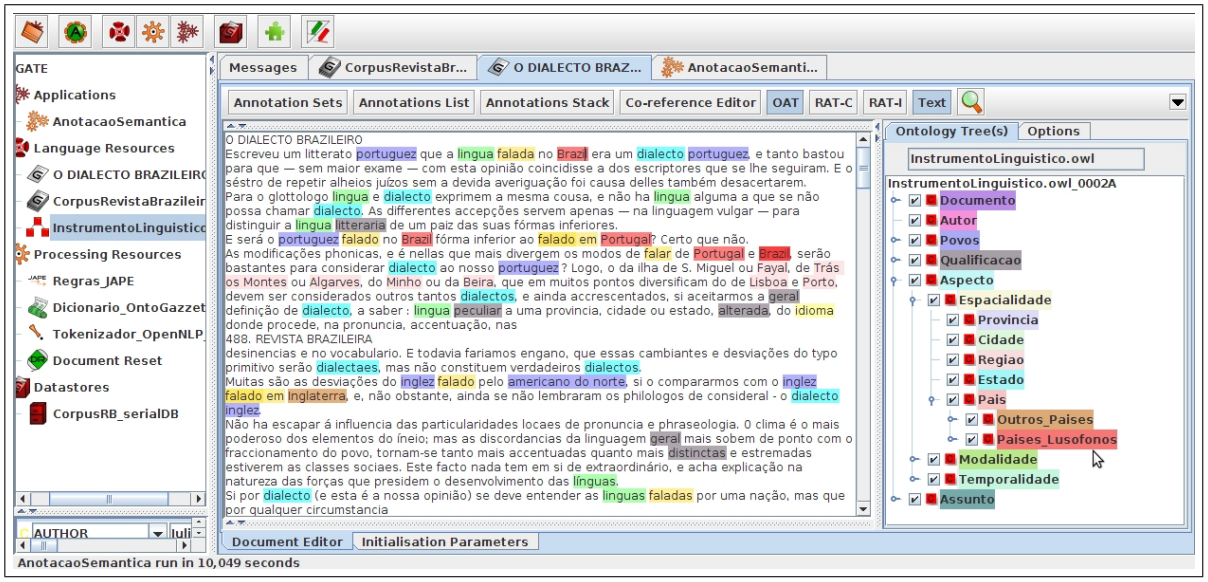


Figura 5.34: Anotações do tipo *Mention* vinculadas à Ontologia InstrumentoLinguistico.

Outra maneira de anotar semanticamente os conceitos primitivos, é através do refinamento do tipo *Mention* em um tipo mais específico, isto é, atribuir a cada menção uma anotação com um tipo que corresponde ao nome da classe a que pertence. Para isso, são construídas regras que seguem o modelo abstrato representado na Figura 5.20.

Na Figura 5.35, é possível observar as ocorrências de anotação do tipo *Lingua*, anotadas pela regra **ClasseLingua**. Esse tipo de anotação é realizada em todas as ocorrências localizadas e marcadas com o tipo *Mention* e propriedade *class* "Lingua", que correspondem a uma menção da classe "Lingua" na ontologia. O cursor ao lado direito da figura demonstra que a anotação do tipo "Lingua" foi selecionada para visualização.

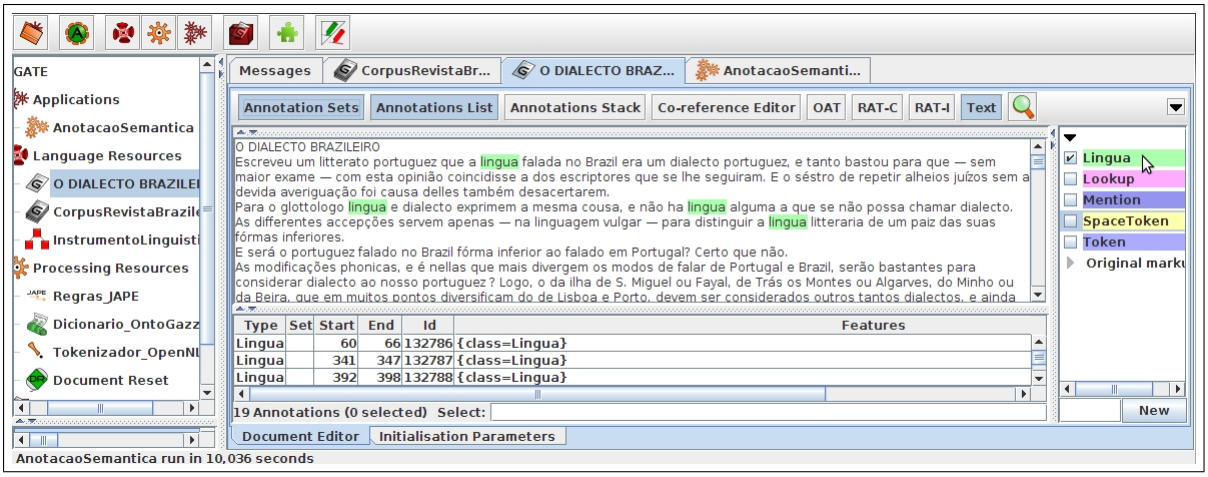


Figura 5.35: Anotações do tipo *Lingua* nas ocorrências localizadas correspondentes a uma menção da classe "Lingua".

Esse tipo de anotação pode ser realizada para todas as classes, ou apenas para as que contribuem na anotação de conceitos derivados mais complexos. Os próximos exemplos vão demonstrar como são realizadas as anotações semânticas dos Conceitos Derivados.

Anotação Semântica dos Conceitos Derivados

A anotação dos conceitos derivados é a marcação com tipos mais expressivos que indicam os relacionamentos semânticos extraídos da ontologia de domínio entre conceitos primitivos e conceitos derivados. Para isso, são construídas regras que podem ser do tipo Relacionamento SuperClasse e SubClasse, Relacionamento Classe e Propriedade ou Relacionamento Classes e Conceitos Derivados, apresentadas na Seção 5.3.2.

Essas regras atribuem anotações de tipos variados de acordo com o determinado por cada regra. A seguir, são apresentados alguns exemplos que demonstram a execução de cada tipo de regra.

Nas regras do tipo Relacionamento SuperClasse e SubClasse, representada como uma abstração na Figura 5.22, subclasses são relacionadas à superclasse e recebem uma marcação do tipo *NomeSuperClasse*. Um exemplo desse tipo é a regra **SuperClasseAssunto**, exibida na Figura 5.23.

O resultado da execução dessa regra pode ser visualizado na Figura 5.36, no centro da figura observam-se as ocorrências marcadas com o tipo Assunto, selecionado ao lado direito para exibição. O cursor seleciona a ocorrência "lingua" permitindo a visualização dos detalhes da anotação do tipo *Assunto*. Os valores para as propriedades de anotação subclasse e superclasse são "Lingua" e "Assunto", respectivamente. Abaixo é possível observar a listagem das

anotações, a linha selecionada mostra os detalhes da ocorrência selecionada. Ao todo, foram feitas trinta e uma (31) anotações do tipo *Assunto*.

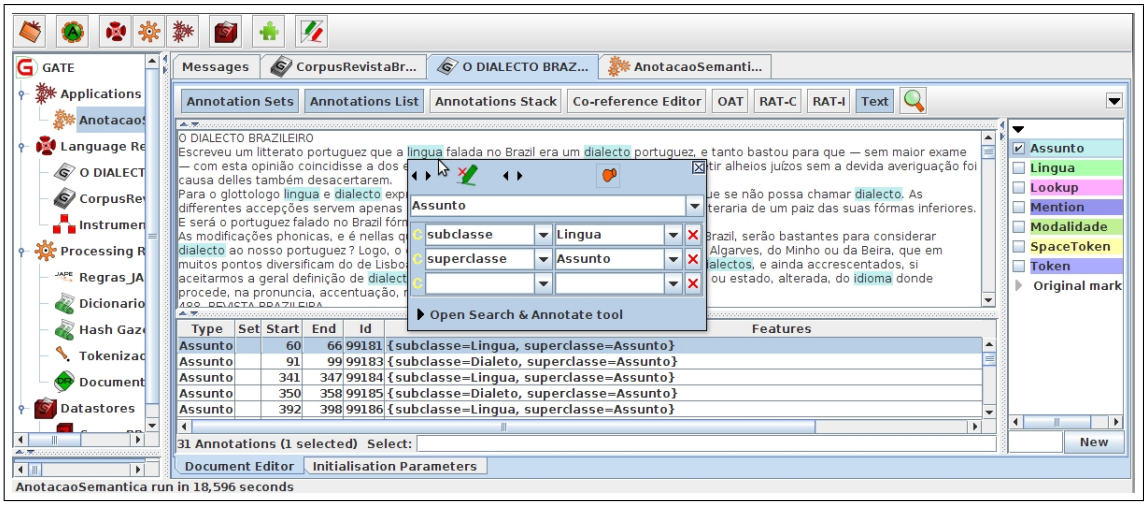


Figura 5.36: Anotações do tipo *Assunto* nas ocorrências de subclasses localizadas.

As anotações semânticas, de conceitos primitivos e derivados, apresentadas até agora foram feitas por regras que são generalizadas em seus respectivos modelos abstratos. Essas regras são independentes de domínio e facilmente adaptadas para outro contexto. Na sequência, são apresentados exemplos com os resultados da execução de regras de definem conceitos derivados mais complexos, nesse sentido, sua construção é totalmente dependente do domínio e exige trabalho em conjunto com o especialista de domínio.

Para o tipo de regra Relacionamento Classe e Propriedade é apresentado um exemplo na Figura 5.37, onde é possível observar os resultados da execução da regra **Naturalidade_portuguez**, que foi construída de maneira análoga à regra "Naturalidade_brazileiro" representada pela Figura 5.24.

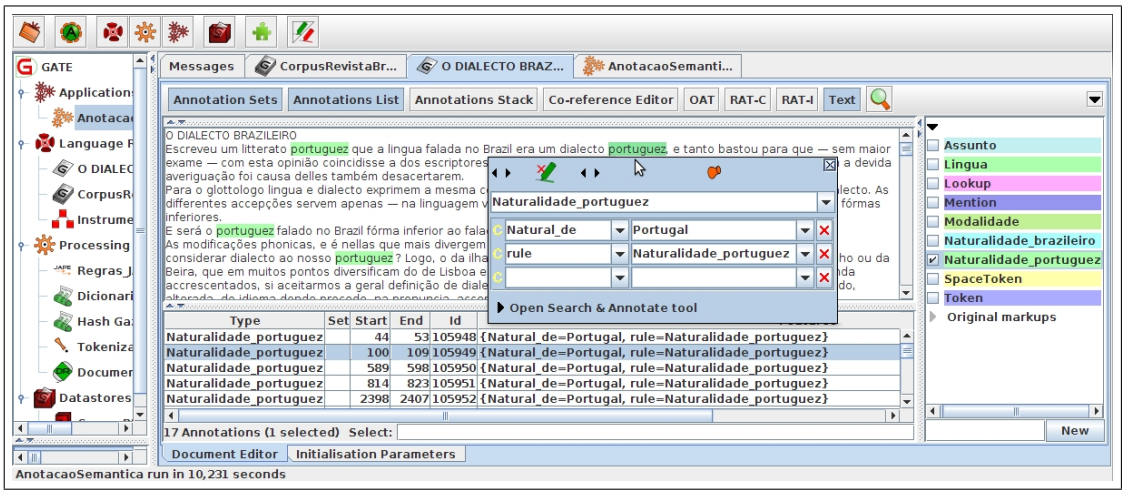


Figura 5.37: Anotações do tipo *Naturalidade_portuguez* nas ocorrências localizadas.

Na Figura 5.37 é possível visualizar as anotações do tipo "Naturalidade_portuguez" nas ocorrências localizadas. O detalhamento da anotação na segunda ocorrência no documento "O Dialecto Brazileiro", mostra que a propriedade Natural_de tem valor "Portugal" e a propriedade *rule* valor "Naturalidade_portuguez". Com a regra foram realizadas dezessete anotações desse tipo no documento.

Um exemplo de regra do tipo Relacionamento Classes e Conceitos Derivados, é apresentado na Figura 5.38, onde é possível observar os resultados da execução da regra **Dialeto_portuguez**, que foi construída de maneira análoga à regra "Dialeto_brazileiro" representada pela Figura 5.24.

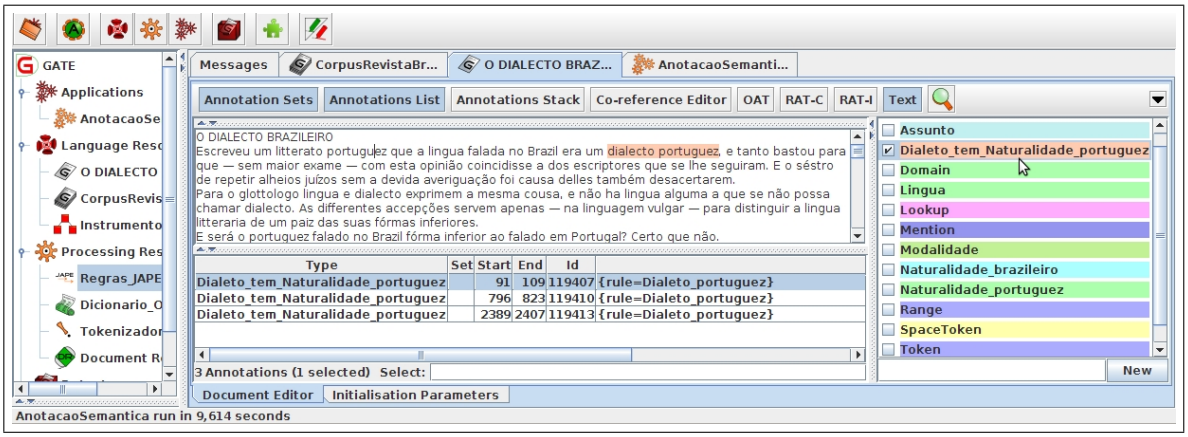


Figura 5.38: Anotações do tipo *Dialeto_portuguez* nas ocorrências localizadas.

Na Figura 5.38 é possível observar na região central, a ocorrência destacada com a cor laranja que representa a anotação do tipo "Dialeto_tem_Naturalidade_portuguez".

Por fim, um exemplo de anotação semântica em conceitos derivados mais complexos pode ser observado na Figura 5.39

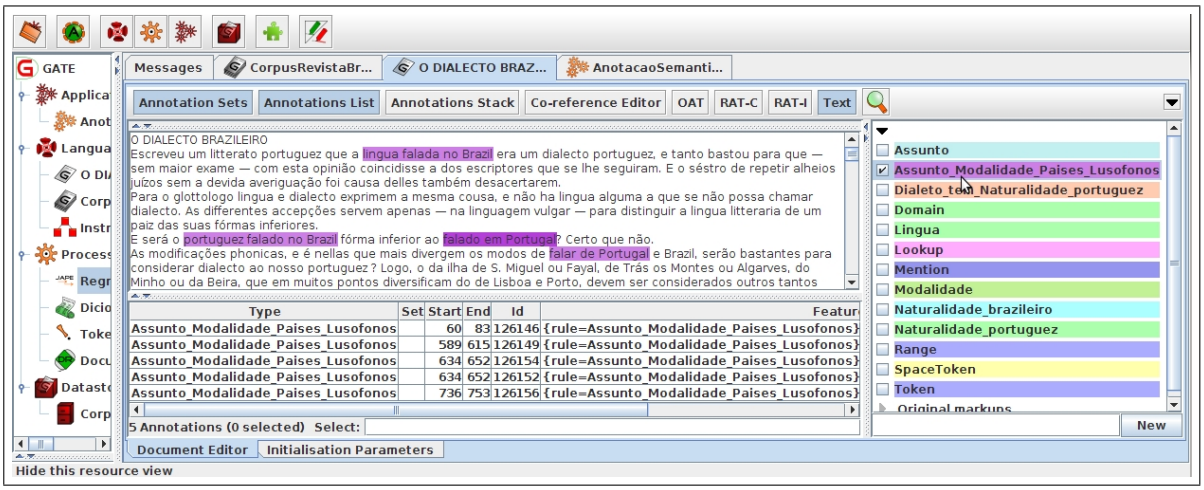


Figura 5.39: Anotações do tipo *Assunto_Modalidade_Paises_Lusofonos* nas ocorrências localizadas.

Na Figura 5.39, o exemplo de regra utilizado anota diferentes conceitos derivados correlacionados, como nos resultados da execução da regra **Assunto_Modalidade_Paises_Lusofonos**. Essa regra, representada na Figura 5.26, busca por um padrão estabelecido junto ao especialista de domínio, capaz de anotar diferentes conceitos derivados. Nessa caso, os conceitos "lingua falada no Brazil", "portuguez falado no Brazil", "falado em Portugal" e "falar de Portugal", foram os conceitos derivados localizados que atendem a regra em questão.

É importante salientar que os exemplos de anotação foram demonstrados individualmente, porém, podem ser visualizados em conjunto para auxiliar na análise dos dados. Por exemplo, na Figura 5.40, pode ser visualizado o resultado da seleção dos tipos: "Assunto", "Assunto_Modalidade_Paises_Lusofonos", "Dialeto_tem_Naturalidade_portuguez", "Lingua" e "Naturalidade_Portuguez".

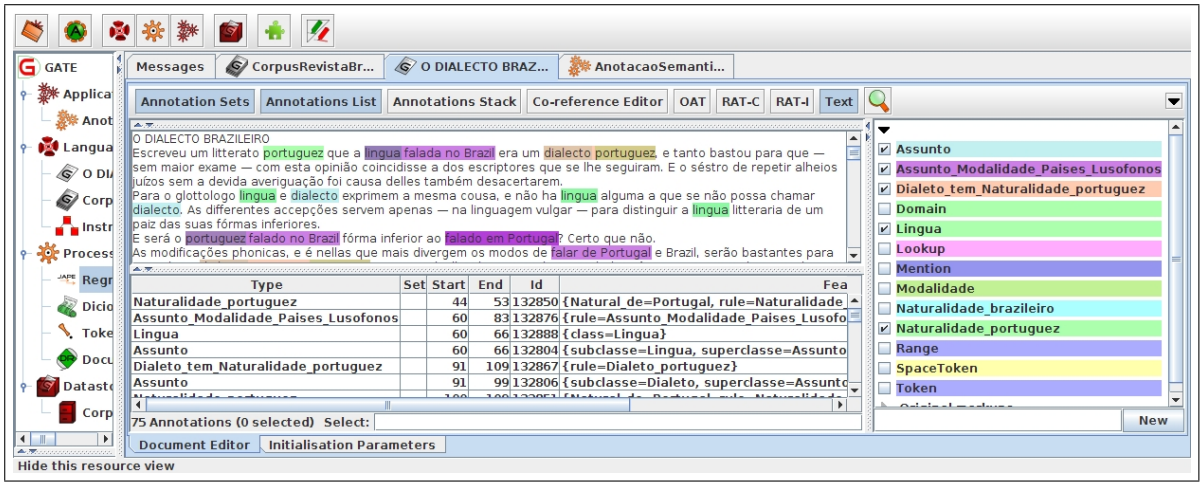


Figura 5.40: Vários tipos de Anotações selecionadas para visualização das ocorrências localizadas.

O resultado da etapa de Anotação Semântica é um córpus com documentos no formato XML Semântico. Um exemplo do documento XML pode ser observado na Figura 5.41.

O_DIALECTO_BRAZILEIRO_AllAnot_Format.xml

```
1 O DIALECTO BRAZILEIRO
2 Escreveu um litterato <Naturalidade_portuguez>portuguez</Naturalidade_portuguez> que a
  <Assunto_Modalidade_Paises_Lusofonos><Assunto><Lingua>lingua</Lingua></Assunto> <Modalidade>falada</
  Modalidade> no Brazil</Assunto_Modalidade_Paises_Lusofonos> era um
  <Dialeto_tem_Naturalidade_portuguez><Assunto>dialecto</Assunto> <Naturalidade_portuguez>portuguez</
  Naturalidade_portuguez></Dialeto_tem_Naturalidade_portuguez>, e tanto bastou para que — sem maior exame —
  com esta opinião coincidisse a dos escriptores que se lhe seguiram. E o séstro de repetir alheios juízos
  sem a devida averiguação foi causa delles também desacertarem.
3 Para o glottologo <Assunto><Lingua>lingua</Lingua></Assunto> e <Assunto>dialecto</Assunto> exprimem a
  mesma cousa, e não ha <Assunto><Lingua>lingua</Lingua></Assunto> alguma a que se não possa chamar
  <Assunto>dialecto</Assunto>. As differentes accepções servem apenas — na linguagem vulgar — para
  distinguir a <Assunto><Lingua>lingua</Lingua></Assunto> litteraria de um paiz das suas fórmas inferiores.
4 E será o <Assunto_Modalidade_Paises_Lusofonos><Naturalidade_portuguez>portuguez</Naturalidade_portuguez>
  <Modalidade>falado</Modalidade> no Brazil</Assunto_Modalidade_Paises_Lusofonos> fórma inferior ao
  <Assunto_Modalidade_Paises_Lusofonos><Assunto_Modalidade_Paises_Lusofonos><Modalidade><Modalidade>falado</
  Modalidade> em</Modalidade> Portugal</Assunto_Modalidade_Paises_Lusofonos></
  Assunto_Modalidade_Paises_Lusofonos>? Certo que não.
```

XML ▾ Largura das tabulações: 8 ▾ Lin 5, Col 450 INS

Figura 5.41: Documento XML Semântico com anotações dos tipos definidos por regras em um trecho do documento "O Dialecto Brazileiro".

Na Figura 5.41, pode ser visualizada a representação do documento XML Semântico com as anotações dos tipos definidos por regras para os conceitos primitivos e os conceitos derivados, em um trecho do documento "O Dialecto Brazileiro". As anotações estão representadas por marcações de acordo com o tipo correspondente, $<$ *Tipo*$>$ **ocorrência** $</$ *Tipo*$>$, em cada ocorrência localizada. Conforme a figura, observa-se que as anotações podem se sobrepor.

No próximo capítulo, são discutidos os resultados e a avaliação da abordagem desenvolvida.

Capítulo 6

DISCUSSÃO DOS RESULTADOS E AVALIAÇÃO DA PROPOSTA

6.1 Documentos Analisados

O córpus formado por cinco documentos foi submetido ao processo de anotação semântica automática apresentado no Capítulo 5, os detalhes de cada documento desse córpus são apresentados na Tabela 4.1. Na Tabela 6.1, os documentos estão identificados por letras ao lado esquerdo e os respectivos números de *tokens* identificados estão ilustrados ao lado direito.

Tabela 6.1: Córpus Revista Brazileira.

	Documentos Históricos	**Autor**	***Tokens***
A	Estudos Lexicographicos Do Dialecto Brazileiro IV	Macedo Soares	13502
B	O Dialecto Brazileiro	Pacheco Junior	3449
C	Estudos Lexicographicos Do Dialecto Brazileiro VII	Macedo Soares	3330
D	Uma Questão Glottologica	Pacheco Junior	3802
E	Questões de Linguistica	Paranhos da Silva	3185
		Número total de *Tokens*	**27268**

Foram identificados um total de 27268 *tokens* nos documentos. A seguir são apresentados os resultados das anotações semânticas realizadas pelo especialista de domínio no *Gold Standard*.

6.2 *Gold Standard*

O *Gold Standard* utilizado para avaliação da proposta é realizado manualmente pelo especialista de domínio. Foi desenvolvido a partir do consenso determinado pelo especialista na Seção

4.1 e da análise da Ontologia InstrumentoLinguistico, apresentada na Seção 4.2. Esse instrumento pode ser utilizado para medição de desempenho das anotações realizadas pela abordagem de anotação semântica automática, conforme apontado por (MAYNARD; PETERS; LI, 2006) e (ROBERTS et al., 2009).

Para construir esse *Gold Standard*, o especialista de domínio realizou uma anotação de acordo com os tipos criados pela abordagem. Isto é, o especialista usa o mesmo *Gold Standard* final resultante do consenso, atribuindo para as marcações destacadas anteriormente, algum tipo específico que melhor representa o conceito primitivo ou o conceito derivado, ou ambos. Os tipos são predeterminados e o especialista escolhe entre esses tipos algum para atribuir a cada *token*, aquele que melhor o representa.

Na Tabela 6.2, são apresentados os tipos de anotação com os respectivos resultados para cada documento do *Gold Standard*.

Tabela 6.2: Resultados no *Gold Standard* para alguns tipos de anotação de Conceitos Derivados.

	Documentos					
Tipo de Anotação	**A**	**B**	**C**	**D**	**E**	**Total**
Assunto	190	66	28	25	114	423
Assunto_temQualificacao	11	6	4	2	13	36
Assunto_temTemporalidadeImprecisa	3	1	0	2	22	28
Dialeto_tem_Naturalidade_brazileiro	6	0	0	0	0	6
Dialeto_tem_Naturalidade_portuguez	1	2	0	0	1	4
Assunto_Lusofonos	20	9	10	1	6	46
Assunto_Modalidade_Paises_Lusofonos	3	4	0	1	2	10

6.3 Resultados da Anotação Semântica Automática

Nesta seção são apresentados os resultados obtidos para cada tipo de anotação em cada um dos documentos. Podem ser observados na Tabela 6.3, os resultados dos tipos de anotação a partir do dicionário *(Lookup)* e a partir das regras que mapeiam os Conceitos Primitivos para a Ontologia InstrumentoLinguistico *(Mention)*.

Esses dados estão sendo apresentados para demonstrar o volume total das anotações baseadas na ontologia. Como pode ser observado, essas anotações são numerosas e genéricas pois não atribuem a cada ocorrência a marcação semântica adequada.

Por isso essas ocorrências recebem, por meio de regras, marcações semânticas que expressam o conhecimento estruturado pela ontologia de domínio. Na Tabela 6.4, podem ser observados os resultados para os tipos de anotações em ocorrências de Conceitos Primitivos, Conceitos

Derivados e Conceitos Derivados mais complexos.

Tabela 6.3: Resultados para os tipos de anotação *Lookup* e *Mention* em cada Documento.

	Documentos					
Tipo de Anotação	**A**	**B**	**C**	**D**	**E**	**Total**
Lookup	569	204	102	114	399	1388
Mention	569	204	102	114	399	1388

Tabela 6.4: Resultados para os tipos de anotação de Conceitos Primitivos e Conceitos Derivados em cada Documento.

	Documentos					
Tipo de Anotação	**A**	**B**	**C**	**D**	**E**	**Total**
Dialeto	15	11	1	3	34	64
Idioma	7	1	2	1	9	20
Lingua	31	19	4	13	54	121
Outros_Assuntos	142	35	23	9	17	226
Assunto	195	66	30	26	114	431
Naturalidade_portuguez	58	17	13	11	48	147
Naturalidade_brazileiro	17	3	6	2	23	51
Assunto_temQualificacao	11	5	4	0	14	34
Assunto_temTemporalidadeImprecisa	3	1	0	2	24	30
Dialeto_tem_Naturalidade_brazileiro	6	0	0	0	0	6
Dialeto_tem_Naturalidade_portuguez	1	3	0	0	1	5
Assunto_Lusofonos	19	9	7	1	7	43
Assunto_Modalidade_Paises_Lusofonos	4	5	0	1	2	12

O tipo de anotação **Assunto** é obtido pela regra SuperClasseAssunto (ver Figura 5.36), que localiza as ocorrências das classes Dialeto, Idioma, Lingua e Outros_Assuntos. Conforme a Tabela 6.4, o tipo foi marcado em 431 ocorrências em todo o córpus. Essas ocorrências representam os termos chave do domínio investigado que são, em geral, genéricas e em grande número. Esse número pode ser reduzido com a aplicação de regras que permitem a manipulação de contextos mais específicos do domínio.

Por exemplo, as regras que localizam:

- **Conceitos Derivados:**
  - Assunto_temQualificacao - 34 ocorrências.
  - Assunto_TemporalidadeImprecisa - 30 ocorrências.
- **Conceitos Derivados mais complexos:**
  - Assunto_Lusofonos - 43 ocorrências.

– Assunto_Modalidade_Paises_Lusofonos - 12 ocorrências.

O tipo Assunto_Lusofonos é o tipo que permite localizar várias ocorrências de diferentes conceitos derivados do domínio estudado. Por se tratar de um tipo mais genérico, além das ocorrências exclusivamente marcadas pelo tipo, podem ser localizadas algumas ocorrências que também são localizadas por outros tipos mais específicos. Por exemplo as ocorrências de "Dialecto_Portuguez", que também são marcadas pelo tipo "Dialeto_tem_Naturalidade_portuguez".

O tipo Assunto_Modalidade_Paises_Lusofonos, é outro exemplo de especialização de algumas das ocorrências marcadas com o tipo Assunto_Lusofonos, por exemplo, "lingua falada no Brasil". No entanto, com esse tipo mais especializado são localizadas ocorrências únicas não marcadas anteriormente como "falado em Portugal".

É fundamental determinar tipos mais abrangentes e mais específicos para garantir que as ocorrências sejam localizadas e marcadas adequadamente e auxiliem em um futuro processo de recuperação de informação que utilize as anotações semânticas. Por exemplo, de acordo com a Tabela 6.4, no documento identificado pela letra "**C**" ocorrem 7(sete) ocorrências do tipo Assunto_Lusofonos e 0 (zero) ocorrências para os outros tipos que determinam o domínio, que são Assunto_TemporalidadeImprecisa, Dialeto_tem_Naturalidade_brazileiro, Dialeto_tem_Naturalidade_portuguez, Assunto_Modalidade_Paises_Lusofonos. Nesse caso, é imprescindível a anotação do tipo Assunto_Lusofonos para localizar ocorrências nesse documento.

6.4 Comparação com o *Gold Standard*

As comparações com o *Gold Standard* tem por objetivo analisar o desempenho da abordagem de anotação semântica automática. As análises são feitas sobre um tipo de anotação para conceito derivado e seis tipos de anotação para conceitos derivados mais complexos.

Na Tabela 6.5, pode ser visualizado o resultado gerado pelo *Corpus Quality Assurance* no *GATE*. A linha exibida representa o tipo de anotação Assunto, as colunas representam respectivamente: Precisão (P), Revocação (R) e *F1-Score* (F1), considerando o critério *Strict* e considerando o critério *Lenient*.

Tabela 6.5: Comparação com o *Gold Standard* - tipo Assunto.

	Strict			***Lenient***		
Tipo de Anotação	P	R	F1	P	R	F1
Assunto	0.98	1.00	0.99	0.98	1.00	0.99

Na Tabela 6.6, podem ser visualizados os resultados gerados pelo *Corpus Quality Assurance* no GATE. As linhas exibidas representam os tipos de anotação de Conceitos Derivados mais complexos. As duas últimas linhas representam respectivamente o *Macro* e o *Micro summary*. *Micro summary* calcula as medidas de Precisão, Revocação e *F-Measure* de maneira geral, pois trata o córpus como um grande documento. No *Macro summary* as medidas são calculadas para todos os tipos de anotação, em seguida é calculada a média dos resultados (CUNNINGHAM et al., 2014).

Tabela 6.6: Comparação das anotações no córpus com as do *Gold Standard* - tipos Conceitos Derivados mais complexos.

	Strict			***Lenient***		
Tipo de Anotação	P	R	F1	P	R	F1
Assunto_temQualificacao	0.97	0.92	0.94	0.97	0.92	0.94
Assunto_temTemporalidadeImprecisa	0.90	0.96	0.93	0.93	1.00	0.97
Dialeto_tem_Naturalidade_brazileiro	1.00	1.00	1.00	1.00	1.00	1.00
Dialeto_tem_Naturalidade_portuguez	0.80	1.00	0.89	0.80	1.00	0.89
Assunto_Lusofonos	0.93	0.87	0.90	0.95	0.89	0.92
Assunto_Modalidade_Paises_Lusofonos	0.82	0.90	0.86	0.91	1.00	0.95
Macro summary	0.90	0.94	0.92	0.93	0.97	0.94
Micro summary	0.92	0.92	0.92	0.95	0.94	0.94

Na avaliação os resultados utilizando o Strict e o Lenient variam de 0.80 a 1.00. Mesmo usando o Strict observa-se um alto grau de coincidência entre o *gold standard* e o córpus. De maneira geral, os resultados obtidos foram altos.

Observa-se que o conceito derivado complexo Dialeto_tem_Naturalidade_brazileiro possui 100% de coincidência. Uma informação interessante é que esse tipo apresenta poucas ocorrências, mesmo sendo um conceito chave da discussão da constituição da Língua Portuguesa no Brasil.

Análise por Tipo de Anotação nos Documentos

Na Figura 6.1, podem ser observados os resultados obtidos por meio da ferramenta *Annotation Diff*. Esses resultados mostram a diferença entre as anotações do *Gold Standard* e as anotações automáticas no documento **Questões de Linguistica** para o tipo Assuntos_Lusofonos.

Na parte superior da interface, Figura 6.1, podem ser escolhidos o documento, o conjunto de anotação e o tipo de anotação, além do botão *Compare*. A parte central está dividida em duas partes, na parte esquerda estão os dados do *Gold Standard (Key)*, na parte direita os dados da anotação automática *(Response)*. Por fim, na parte inferior estão os medidas de desempenho da anotação automática.

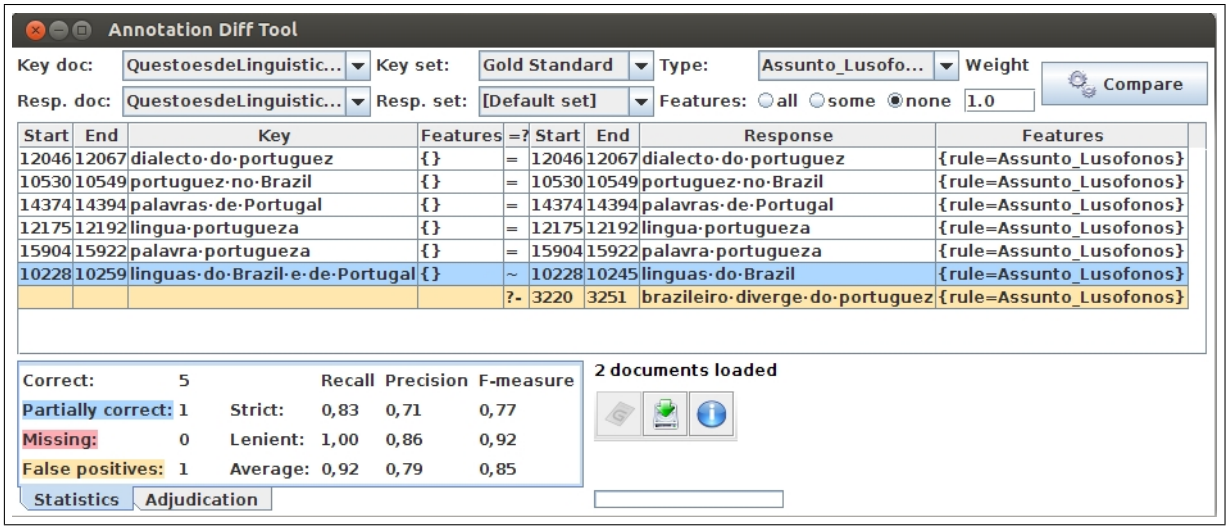


Figura 6.1: Diferença entre o *Gold Standard* e o automático, no documento Questões de Linguistica e tipo Assuntos_Lusofonos.

Por exemplo, a ferramenta identificou cinco ocorrências corretas marcadas em branco na parte central. Uma ocorrência parcialmente correta marcada em azul e uma ocorrência de falso positivo marcada em amarelo. A ocorrência "linguas do Brazil e de Portugal" anotada pelo *Gold Standard* foi parcialmente anotada pela anotação automática como "linguas do Brazil". A ocorrência "brazileiro diverge do portuguez" é marcada somente pela anotação automática, caracterizando um falso positivo pois ela toma dois conceitos primitivos isolados como uma ocorrência do tipo Assuntos_Lusofonos. Nesse caso, é possível verificar uma diferença maior entre as medições ao utilizar o *Strict* e o *Lenient*.

Nos casos em que a ocorrência foi perdida, isto é, anotada unicamente pelo *Gold Standard*, a ferramenta sinaliza a marcação em vermelho *(Missing)*. Por exemplo, no documento *Estudos Lexicographicos Do Dialecto Brazileiro IV* as ocorrências "termos correntes do Brazil", "t. port" e "termo popular brazileiro" foram marcadas somente no *Gold Standard*, ver Figura 6.2.

O tipo Assuntos_Lusofonos não foi marcado nas ocorrências em que há Qualificacao do Assunto com Naturalidade_brazileiro ou com Paises_Lusofonos Brazil. Na verdade não foi criada uma regra mais específica para localizar esse tipo de ocorrência. Provavelmente, o especialista de domínio atribuiu às ocorrências esse tipo por falta de um tipo mais adequado, por exemplo, Assunto_temQualificacao_Lusofonos. Nesse caso, uma solução seria a construção da regra mais específica.

No documento *O Dialecto Brazileiro*, a ocorrência marcada como parcialmente correta "modos de falar de Portugal" , ao invés de "modos de falar de Portugal e Brazil", para o tipo Assunto_Modalidade_Paises_Lusofonos é exemplo de casos em que o recurso linguístico da

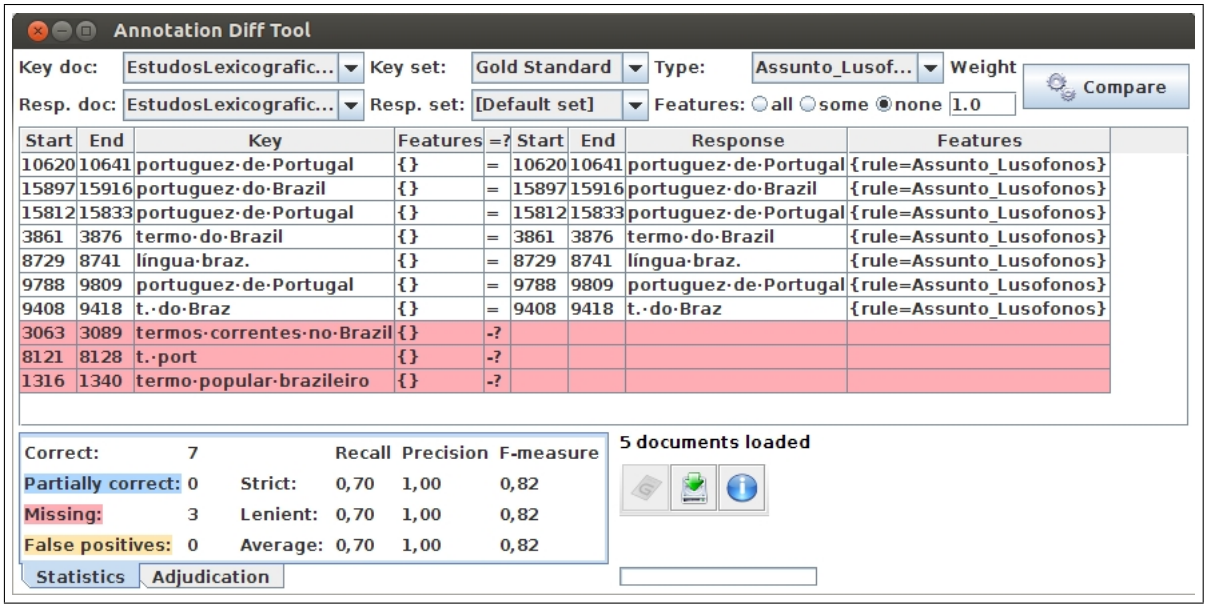


Figura 6.2: Diferença entre o *Gold Standard* e a anotação automática no documento Estudos Lexicographicos Do Dialecto Brazileiro IV e tipo Assuntos_Lusofonos.

elipse diferencia as anotações. Como visto em outros casos a anotação automática não localiza os termos que estão sendo subentendidos ou ocultos na frase.

Como parcialmente correto é marcada ainda a ocorrência "do portuguez antigo", ao invés de "do portuguez antigo para o moderno", com o tipo Assunto_TemporalidadeImprecisa. Apesar de ser uma informação interessante, a regra não localiza a passagem do tempo mas apenas momentos isolados. O falso positivo "dialecto ao nosso portuguez" também atribui a dois conceitos primitivos isolados uma ocorrência do tipo Dialeto_tem_Naturalidade_portuguez, limitação já identificada anteriormente.

A confusão entre a abreviação (t) da palavra "termo" e a letra (t) para representar um fonema ou a letra do alfabeto, gera falsos positivos na anotação automática. No documento *Uma Questão Glottologica* verificamos uma ocorrência, no documento *Estudos Lexicographicos Do Dialecto Brazileiro VII* duas ocorrências e no documento *Estudos Lexicographicos Do Dialecto Brazileiro IV* cinco ocorrências.

Conforme mencionado anteriormente, a construção de novas regras pode ajudar a solucionar as limitações.

6.5 Análise em uma amostra de Revistas Completas

Para a realização desse teste ampliamos a amostra para nove Revistas Brazileiras completas do período de 1879-1900 que inclui documentos da fase Midosi e José Veríssimo.

Foram localizados o total de 1.332.720 ocorrências de *tokens* nas Revistas Brazileiras. Dentre as quais foi possível localizar ocorrências de conceitos primitivos e conceitos derivados do domínio estudado nesta dissertação.

A seguir estão identificados os tipos de anotação realizadas pela abordagem automática nas Revistas Brazileiras.

Tipos de Anotação:

1. Dialeto.
2. Idioma.
3. Lingua.
4. Assunto.
5. Assunto_temQualificacao.
6. Assunto_temTemporalidadeImprecisa.
7. Dialeto_tem_Naturalidade_brazileiro.
8. Dialeto_tem_Naturalidade_portuguez.
9. Assunto_Lusofonos.
10. Assunto_Modalidade_Paises_Lusofonos.

Na Tabela 6.7, são apresentados os resultados do número de ocorrências localizadas em cada revista.

Como pode ser observado, as ocorrências de conceitos primitivos (tipos 1-Dialeto, 2-Idioma e 3-Lingua) aparecem em grande quantidade conforme expectativa do especialista de domínio. Isto se deve ao fato de que os termos chave, representados na totalidade pela superclasse Assunto (tipo 4), podem ser conceitos primitivos isolados que não se relacionam com nenhuma outra ocorrência extraída da ontologia.

Para os conceitos derivados esses números são reduzidos, obviamente porque quanto mais complexo o conceito derivado a ser localizado por uma regra, mais expressiva será a ocorrência em relação ao domínio discutido e mais precisa será a caracterização do documento.

A abordagem de anotação semântica automática identificou corretamente todos os documentos que foram utilizados para a construção do *Gold Standard*. Além disso, também foi

Tabela 6.7: Resultados para alguns tipos de anotação em cada Revista Brazileira da amostra.

	Tipos de Anotação									
Revista Brazileira	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RB_1879_0001.pdf	7	8	91	371	1	4	0	0	11	1
RB_1879_0002.pdf	2	12	99	328	3	1	0	0	6	0
RB_1880_0003.pdf	5	6	60	260	8	0	4	0	16	0
RB_1880_0004.pdf	8	12	64	358	9	1	1	0	12	4
RB_1880_0005.pdf	13	12	93	323	7	2	2	3	27	5
RB_1881_0001_09.pdf	7	6	62	237	5	5	0	1	16	0
RB_1881_0007.pdf	31	7	80	281	13	17	2	1	9	2
RB_1881_0008.pdf	3	7	69	261	10	1	0	0	18	1
RB_1895_0001.pdf	0	12	39	197	2	1	0	0	4	2
Total	76	82	657	2616	58	32	9	5	119	15

localizado corretamente o documento *Estudos de Linguistica* de Said Ali, sinalizado anteriormente por Bechara (2005) e Gonçalves (2012), porém não utilizado no córpus desta pesquisa.

Através desse teste foi possível ainda localizar um documento não identificado anteriormente, nem por Bechara (2005), nem por Gonçalves (2012). O documento *A Poesia Popular no Brazil* de Sylvio Roméro, localizado por ocorrências de conceitos derivados como "portuguez do Brazil" e "portuguez brazileiro", discute a problemática envolvida nesta dissertação e que não foi mencionado pelos autores citados, confirmando a eficácia da abordagem automática na localização e marcação de elementos que descrevem a conceitualização do domínio discutido.

Vale lembrar que essas revistas são os arquivos originais disponibilizados pela Hemeroteca Digital Brasileira[1] e não receberam nenhum tratamento nos erros de OCR, portanto estão sujeitas a maior nível de ruído no processo de anotação automática. Um exemplo ocorre no documento *Uma Questão Glottologica* onde a ocorrência "portuguez fa-lado no Brazil", originalmente separada por hífen, só é marcada pelo tipo Assunto_Lusofonos, sendo que deveria ter sido marcada também com o tipo Assunto_Modalidade_ Paises_Lusofonos.

Mais uma vez foi possível comprovar que a informação relevante do tipo "Dialeto_tem_ Naturalidade_brazileiro", apresenta poucas ocorrências, mesmo considerando uma amostra maior como o conjunto de Revistas Brazileiras.

[1] http://hemerotecadigital.bn.br/

Capítulo 7

Conclusão e Trabalhos Futuros

Nesta dissertação apresentamos uma abordagem para anotação semântica automática baseada em ontologia para o estudo do Português Brasileiro em documentos históricos do final do século XIX.

Este trabalho adotou a construção de listas derivadas da ontologia, o dicionário, por isso foi possível localizar as ocorrências de conceitos primitivos no córpus sem uma análise morfossintática. Por meio das listas também resolvemos os problemas relacionados à variação de grafia comumente encontrados em documentos históricos.

Com regras derivadas da Ontologia InstrumentoLinguistico, foi possível realizar a anotação semântica nos documentos históricos localizando ocorrências relacionadas ao domínio investigado, mais especificamente a constituição da Língua Portuguesa no Brasil.

Os resultados encontrados a partir de análise comparativa entre o *Gold Standard* e as anotações provenientes da abordagem automática apresentam altos índices de coincidência, comprovando que o processo é eficiente e que a Ontologia InstrumentoLinguistico define adequadamente o domínio discutido.

Os resultados mostraram ainda, que é preciso construir regras mais específicas em função do domínio para uma identificação mais precisa de documentos que possuem ocorrências de conceitos derivados complexos. Um resultado bastante interessante, do ponto de vista do especialista de domínio, foi a identificação de documentos desse domínio na amostra de Revistas Brazileiras completas que não haviam sido identificadas em trabalhos reconhecidos como de Gonçalves (2012) e de Bechara (2005), como apresentado na Seção 6.5.

7.1 Trabalhos futuros

A seguir são destacados os trabalhos futuros:

- Validar a Ontologia InstrumentoLinguistico por outros especialistas de domínio.
- Investir na construção de novas regras para localizar outros conceitos derivados mais complexos.
- Analisar outros documentos do domínio produzidos em épocas distintas, com a abordagem de anotação automática.
- Utilizar múltiplas ontologias na abordagem desenvolvida.
- Aplicar a abordagem de anotação semântica automática em outros domínios de conhecimento.

REFERÊNCIAS

ALUÍSIO, S. *Córpus Históricos, Recursos Léxicos e Ferramentas para a tarefa de criação de dicionários. I Escola Brasileira de Lingüística Computacional USP, Setembro de 2007*. 2007. `http://www.letras.etc.br/ebralc/Aluisio2.pdf`. Último acesso em: 14/02/2014.

AUROUX, S. *A revolução tecnológica da gramatização*. [S.l.]: Unicamp Campinas, 1992.

BECHARA, E. *A língua portuguesa na revista brasileira*. [S.l.]: Academia Brasileira de Letras, 2005.

BERNERS-LEE, T. et al. The semantic web. *Scientific american*, New York, NY, USA:, v. 284, n. 5, p. 28–37, 2001.

BORST, W. N. Construction of engineering ontologies for knowledge sharing and reuse. Universiteit Twente, 1997.

BREWSTER, C. Techniques for automated taxonomy building: Towards ontologies for knowledge management. In: *Proceedings CLUK Research Colloquium*. [S.l.]: Springer Verlag, 2002. p. 27–28.

CÂNDIDO JR, A.; ALUÍSIO, S. M. Procorph: um sistema de apoio à criação de dicionários históricos. In: ACM. *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*. [S.l.], 2008. p. 347–352.

CÂNDIDO JR, A.; ALUÍSIO, S. M. Building a corpus-based historical portuguese dictionary: Challenges and opportunities. *TAL*, v. 50, n. 2, p. 73–102, 2009.

COWIE, J.; LEHNERT, W. Information extraction. *Communications of the ACM*, ACM, v. 39, n. 1, p. 80–91, 1996.

CUNNINGHAM, H. et al. *Developing Language Processing Components with GATE Version 8 (a User Guide)*. 2014. `http://gate.ac.uk/sale/tao/tao.pdf`. Último acesso em: 14/05/2014.

CUNNINGHAM, H.; MAYNARD, D.; TABLAN, V. *JAPE: a Java Annotation Patterns Engine (Second Edition)*. [S.l.], nov. 2000.

CUNNINGHAM, H.; WILKS, Y.; GAIZAUSKAS, R. GATE – a General Architecture for Text Engineering. In: *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*. Copenhagen: [s.n.], 1996.

DALIANIS, H.; VELUPILLAI, S. De-identifying swedish clinical text-refinement of a gold standard and experiments with conditional random fields. *J. Biomedical Semantics*, v. 1, p. 6, 2010.

ERNST-GERLACH, A.; FUHR, N. Retrieval in text collections with historic spelling using linguistic and spelling variants. In: ACM. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. [S.l.], 2007. p. 333–341.

FELDMAN, R.; DAGAN, I.; HIRSH, H. Mining text using keyword distributions. *Journal of Intelligent Information Systems*, Springer, v. 10, n. 3, p. 281–300, 1998.

FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. [S.l.]: Cambridge University Press, 2007.

FRAKES, W. B.; BAEZA-YATES, R. Information retrieval: data structures and algorithms. Prentice Hall PTR, 1992.

GALVES, C.; FARIA, P. *Tycho Brahe Parsed Corpus of Historical Portuguese.* 2010. Http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html.

GIUSTI, R. et al. Automatic detection of spelling variation in historical corpus: An application to build a brazilian portuguese spelling variants dictionary. In: *Corpus Linguistics*. [S.l.: s.n.], 2007.

GONÇALVES, M. R. B. *As teorias lingüísticas da espacialidade : uma agenda dialetológica na gramatização do português do Brasil*. Tese (Doutorado), 2012.

GRUBER, T. Ontology. *Encyclopedia of database systems*, Springer, p. 1963–1965, 2009.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition*, Elsevier, v. 5, n. 2, p. 199–220, 1993.

GUARINO, N. *Formal ontology in information systems: proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*. [S.l.]: Ios PressInc, 1998.

HIROHASHI, A. *Aprendizado de regras de substituição para normatização de textos históricos*. Dissertação (Mestrado) — dissertação (Mestrado em Ciências de Computação e Matemática Computacional), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

IRIA, J. et al. Integrating information extraction, ontology learning and semantic browsing into organizational knowledge processes. 2004.

KEMPKEN, S.; LUTHER, W.; PILZ, T. Comparison of distance measures for historical spelling variants. In: *Artificial Intelligence in Theory and Practice*. [S.l.]: Springer, 2006. p. 295–304.

KIRYAKOV, A. et al. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, v. 2, n. 1, p. 49–79, 2004.

KLYNE, G.; CARROLL, J. J. Resource description framework (rdf): Concepts and abstract syntax. 2006.

LI, Y.; BONTCHEVA, K. Hierarchical, perceptron-like learning for ontology-based information extraction. In: ACM. *Proceedings of the 16th international conference on World Wide Web*. [S.l.], 2007. p. 777–786.

MAKS, I.; VOSSEN, P. Annotation scheme and gold standard for dutch subjective adjectives. In: *LREC*. [S.l.: s.n.], 2010.

MAYNARD, D.; PETERS, W.; LI, Y. Metrics for evaluation of ontology-based information extraction. In: EDINBURGH, UK. *International world wide web conference*. [S.l.], 2006.

MAYNARD, D. et al. Ontology-based information extraction for market monitoring and technology watch. In: *ESWC Workshop "End User Apects of the Semantic Web", Heraklion, Crete*. [S.l.: s.n.], 2005.

MCGUINNESS, D. L.; HARMELEN, F. V. et al. Owl web ontology language overview. *W3C recommendation*, v. 10, n. 2004-03, p. 10, 2004.

MENEGATTI, T. A. *Regras Lingüísticas para Tratamento Computacional da Variação de Grafia e Abreviaturas do Corpus Tycho Brahe.* [S.l.], 2002.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. *Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens*. [S.l.], 2007.

MURAKAWA, C. d. A. A. Lexicografia e história: O dicionário histórico do português do brasil - séculos xvi, xvii, xviii. *Os Estudos Lexicais em Diferentes Perspectivas*, 2009.

NOY, N. F.; MCGUINNESS, D. L. *Ontology Development 101: A Guide to Creating Your First Ontology*. [S.l.], 2001.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. F. *O Processamento automático de textos antigos: Desafios e Experiências*. 2010. `http://humanidadesdigitais.org/publicacoes/`. Último acesso em: 14/02/2014.

PAIXÃO DE SOUSA, M. C.; TRIPPEL, T. Building a historical corpus for classical portuguese: some technological aspects. In: *V International Conference on Language Resources and Evaluation, Genoa: LREC*. [S.l.: s.n.], 2006.

PAN, J. Z. Resource description framework. In: *Handbook on Ontologies*. [S.l.]: Springer, 2009. p. 71–90.

PEREIRA, J. W.; GONÇALVES, M. R. B.; SANTOS, M. T. P. Pré-processamento para recuperação de informação em textos históricos do século XIX. In: SBC. *Proceedings of KDMiLe - Symposium on Knowledge Discovery, Mining and Learning"*. [S.l.], 2013.

POPOV, B. et al. Kim-a semantic platform for information extraction and retrieval. *Natural language engineering*, Cambridge Univ Press, v. 10, n. 3-4, p. 375–392, 2004.

PRUD'HOMMEAUX, E.; SEABORNE, A. et al. Sparql query language for rdf. *W3C recommendation*, v. 15, 2008.

RAYSON, P.; ARCHER, D.; SMITH, N. Vard versus word: A comparison of the ucrel variant detector and modern spellcheckers on english historical corpora. 2005.

ROBERTS, A. et al. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, Elsevier, v. 42, n. 5, p. 950–966, 2009.

ROSS, D. T. Structured analysis (sa): A language for communicating ideas. *Software Engineering, IEEE Transactions on*, IEEE, n. 1, p. 16–34, 1977.

SAGGION, H. et al. Ontology-based information extraction for business intelligence. In: *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*. [S.l.]: Springer, 2007. p. 843–856.

SOUSA, M. P. D.; KEPLER, F. N.; FARIA, P. E-dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. *comunicação apresentada no VIII Encontro de Linguística de Corpus, realizado na UERJ*, v. 13, 2009.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. *Data & knowledge engineering*, Elsevier, v. 25, n. 1, p. 161–197, 1998.

UREN, V. et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: science, services and agents on the World Wide Web*, Elsevier, v. 4, n. 1, p. 14–28, 2006.

USCHOLD, M.; KING, M. Towards a methodology for building ontologies. In: *In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*. [S.l.: s.n.], 1995.

VALE, O. et al. Building a large dictionary of abbreviations for named entity recognition in portuguese historical corpora. In: *The Workshop Programme*. [S.l.: s.n.], 2008. p. 47.

VARGAS-VERA, M. et al. Knowledge extraction by using an ontology-based annotation tool. In: *K-CAP 2001 workshop on Knowledge Markup and Semantic Annotation*. [S.l.: s.n.], 2001.

VELUPILLAI, S. et al. Developing a standard for de-identifying electronic patient records written in swedish: Precision, recall and f-measure in a manual and computerized annotation trial. *International journal of medical informatics*, Elsevier, v. 78, n. 12, p. e19–e26, 2009.

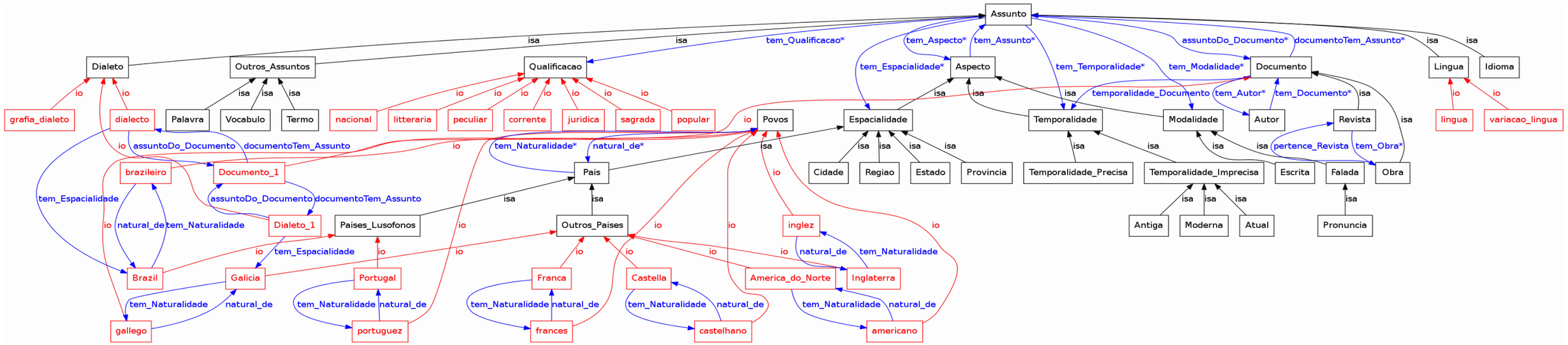
WIMALASURIYA, D. C.; DOU, D. Components for information extraction: ontology-based information extractors and generic platforms. In: ACM. *Proceedings of the 19th ACM international conference on Information and knowledge management*. [S.l.], 2010. p. 9–18.

WIMALASURIYA, D. C.; DOU, D. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, Sage Publications, v. 36, n. 3, p. 306–323, 2010.

YAGUINUMA, C. A.; SANTOS, M. T.; BIAJIZ, M. Meta-ontologia difusa para representação de informações imprecisas em ontologias. In: *Workshop on Ontologies and Metamodeling in Software and Data Engineering*. [S.l.: s.n.], 2007. p. 57–67.

Apendice A

ONTOLOGIA INSTRUMENTOLINGUISTICO


Assunto
Dialeto
Outros_Assuntos
Qualificacao
Aspecto
Documento
Lingua
Idioma
Espacialidade
Temporalidade
Modalidade
Autor
Revista
Povos
Pais
Paises_Lusofonos
Outros_Paises
Cidade
Regiao
Estado
Provincia
Temporalidade_Precisa
Temporalidade_Imprecisa
Escrita
Falada
Obra
Antiga
Moderna
Atual
Pronuncia
Palavra
Vocabulo
Termo
nacional
litteraria
peculiar
corrente
juridica
sagrada
popular
grafia_dialeto
dialecto
brazileiro
Documento_1
Dialeto_1
Brazil
Galicia
Portugal
gallego
portuguez
Franca
Castella
frances
castelhano
inglez
America_do_Norte
Inglaterra
americano
lingua
variacao_lingua
isa
io
tem_Qualificacao*
tem_Aspecto*
tem_Assunto*
tem_Espacialidade*
tem_Temporalidade*
tem_Modalidade*
assuntoDo_Documento*
documentoTem_Assunto*
temporalidade_Documento
tem_Autor*
tem_Documento*
pertence_Revista
tem_Obra*
tem_Naturalidade*
natural_de*
assuntoDo_Documento
documentoTem_Assunto
tem_Espacialidade
tem_Naturalidade
natural_de