



UnB – UNIVERSIDADE DE BRASÍLIA
FCI – Faculdade de Ciência da Informação
PPGCIInf – Programa de Pós-Graduação em Ciência da Informação

AUTO TAVARES DA CAMARA JUNIOR

**PROCESSAMENTO DE LINGUAGEM
NATURAL PARA INDEXAÇÃO
AUTOMÁTICA
SEMÂNTICO-ONTOLÓGICA**

Brasília – DF

2013

AUTO TAVARES DA CAMARA JUNIOR

**PROCESSAMENTO DE LINGUAGEM
NATURAL PARA INDEXAÇÃO
AUTOMÁTICA
SEMÂNTICO-ONTOLÓGICA**

Tese apresentada à banca examinadora como requisito parcial à obtenção do Título de Doutor em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília.

Orientadora: Prof^a. Dr^a. Marisa Bräscher Basílio Medeiros

Brasília – DF

2013

FOLHA DE APROVAÇÃO

Título: Processamento de linguagem natural para indexação automática semântico-ontológica.

Autor: Auto Tavares da Camara Junior

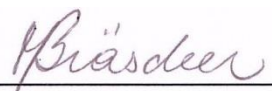
Área de Concentração: Transferência da Informação

Linha de Pesquisa: Arquitetura da Informação

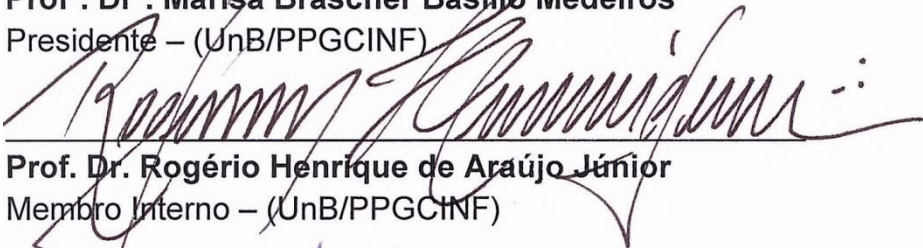
Tese submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Doutor** em Ciência da Informação.

Tese aprovada em: 11 de abril de 2013.

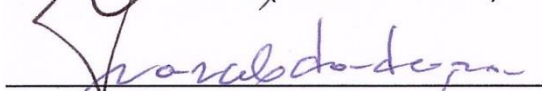
Aprovada por:



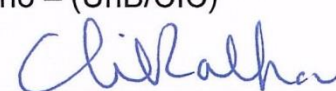
Prof^a. Dr^a. Marisa Bráscher Basílio Medeiros
Presidente – (UnB/PPGCINF)



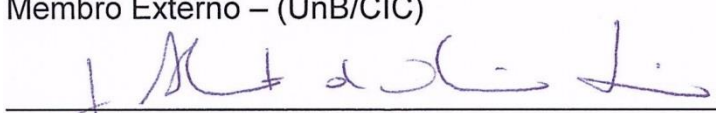
Prof. Dr. Rogério Henrique de Araújo Júnior
Membro Interno – (UnB/PPGCINF)



Prof. Dr. Marcelo Ladeira
Membro Externo – (UnB/CIC)



Prof^a. Dr^a. Célia Ghedini Ralha
Membro Externo – (UnB/CIC)



Prof. Dr. João Alberto de Oliveira Lima
Membro Externo – (Senado Federal)

Prof^a. Dr^a. Dulce Maria Baptista
Suplente – (UnB/PPGCINF)

AGRADECIMENTOS

Primeiramente é preciso reconhecer que é das maiores injustiças reservar um único parágrafo para agradecimento ao orientador. Querida Marisa, esse espaço não é minimamente aceitável para expressar o quanto lhe sou grato por me acompanhar há tantos anos, por ter me resgatado no final da minha pesquisa e por tê-la lapidado, por ter me orientado, por ter suportado minha ansiedade exasperada, enfim, por ter possibilitado que este trabalho fosse concluído. Obrigado, de verdade. Sem você nada disso existiria.

Ao Professor Doutor Jaime Robredo (*in memoriam*), que ao nos deixar criou uma lacuna na Ciência da Informação que jamais será preenchida.

Aos Professores Doutores Rogério Araújo e Marcelo Ladeira, pelas inúmeras contribuições, realinhamentos e melhorias discutidas durante a Qualificação. Esta pesquisa beneficiou-se incomensuravelmente a partir de suas sugestões.

Aos Professores Doutores Rogério Araújo, Marcelo Ladeira, Célia Ghedini, João Lima, Dulce Baptista e Lillian Alvares, pela composição da banca examinadora. É um privilégio ter seu trabalho avaliado por professores e amigos.

Aos meus queridos irmão e irmãs Aparício, Rosemeire, Rosângela, Mírian e Cristiane, que individualmente sempre me inspiraram em alguma fase da minha vida.

A todos aqueles que, direta ou indiretamente, contribuíram para a conclusão desta.

A pesquisa propõe uma arquitetura de indexação automática de documentos utilizando mecanismos de processamento de linguagem natural em nível semântico. Por meio do arranjo de ferramentas e recursos existentes, agregado ao desenvolvimento de *software* para integração, é construído um sistema de indexação automática que utiliza conhecimento modelado em uma ontologia para análise semântica. A aplicação da arquitetura é exemplificada e posta à prova em um conjunto de laudos periciais de crimes cibernéticos produzidos pela Perícia Criminal da Polícia Federal. Os resultados de pesquisa apontam para a melhoria que o aprofundamento à análise semântica do processamento de linguagem natural oferece para a seleção automática de descritores e, por extensão, para a recuperação da informação. Dentre as contribuições inéditas desta tese incluem-se a arquitetura de sistema de informação para indexação automática proposta, a ontologia construída para a análise semântica e as implementações em linguagem de programação Python. Por fim, o trabalho insere-se em uma área de pesquisa que tem sido cada vez mais investigada, no entanto ainda carece de resultados amadurecidos, qual seja o processamento de linguagem natural em língua portuguesa.

Palavras-Chave: Processamento de Linguagem Natural; Indexação Automática; Recuperação de Informação; Ontologia.

ABSTRACT

The research proposes an automatic indexing architecture of documents using natural language processing mechanisms on semantic level. By organizing existing tools and resources, added to software development for integration, an automatic indexing system which uses knowledge modeled by ontology to semantic analysis is built. The applicability of the architecture is exemplified and put into proof on forensics reports of cybercrimes produced by Federal Police Forensics Experts. The research results point to the benefits that semantic analysis on natural language processing offers to automatic descriptor selection and information retrieval. As contributions of this thesis there are the proposed automatic indexing information system architecture, the ontology built to semantic analysis and the implementations on Python programming language. Concluding, the work inserts itself in a research area that has been increasingly more investigated, however still demanding mature results, which is natural language processing on portuguese language.

Keywords: Natural Language Processing; Automatic Indexing; Information Retrieval; Ontology.

LISTA DE FIGURAS

Figura 1 – Tipos de ontologias	23
Figura 2 – Esquema de classificação para usos de ontologias	27
Figura 3 – Estágios de análise em PLN	33
Figura 4 – Estágios retroalimentados de análise em PLN	34
Figura 5 – Modelo plano de análise em PLN	34
Figura 6 – As pontes de Königsberg	70
Figura 7 – Definições de revocação e precisão	82
Figura 8 – Quadro amplo da RI	87
Figura 9 – Fluxograma simplificado do processo de indexação utilizando um tesouro	88
Figura 10 – Relacionamento entre fontes de conhecimento estático	109
Figura 11 – Procedimentos da pesquisa	116
Figura 12 – Proposta de arquitetura de IA	126

LISTA DE QUADROS

Quadro 1 – Distribuição dos laudos de crimes cibernéticos da PF por unidade de registro	103
Quadro 2 – Etiquetas POS para a classificação morfológica	120
Quadro 3 – Laudos de aparelhos celulares com mensagens SMS	121
Quadro 4 – Quantidade de laudos da amostra por temática criminosa	127
Quadro 5 – Cálculo do índice F para consultas submetidas à base da amostra	131

LISTA DE SIGLAS E ABREVIações

AEF	Autômato de estado finito
ASL	Análise semântica latente
BD	Banco de dados
CC	Ciência da computação
CI	Ciência da informação
CKY	Algoritmo de <i>parse</i> sintático Cocke–Kasami–Younger
CP	Código penal
DITEC	Diretoria técnico–científica
FNC	Forma normal de Chomsky
GLC	Gramática livre de contexto
IA	Indexação automática
IM	Indexação manual
INC	Instituto nacional de criminalística
ISL	Indexação semântica latente
KPML	Ambiente de desenvolvimento gramático multilíngue
LC	Linguística computacional

LR	Algoritmo de <i>parse</i> sintático <i>Left–right</i>
MOM	Modelo oculto de Markov
MP	Ministério público
NILC	Núcleo interinstitucional de linguística computacional
NLTK	Ferramental de linguagem natural – <i>Natural language toolkit</i>
NUTEC	Núcleo técnico–científico
OO	Orientação a objetos
OWL	Linguagem ontológica para web – <i>Ontology web language</i>
PCF	Perito criminal federal
PF	Polícia federal
PLN	Processamento de linguagem natural
POS	Categoria morfológica – <i>Part–of–speech</i>
RD	Recuperação de dados
RDF	<i>Framework</i> de descrição de recursos – <i>Resource description framework</i>
RI	Recuperação de informação
SETEC	Setor técnico–científico
SI	Sistema de informação

SMS	Serviço de mensagens curtas – <i>Short message service</i>
SO	Sistema operacional
SQL	Linguagem estruturada de consulta – <i>Structured query language</i>
STJ	Superior tribunal de justiça
TA	Tradução automática
TEF	Transdutor de estado finito
TF-IDF	Frequência do termo–Frequência inversa nos documentos – <i>Term frequency–Inverse document frequency</i>
TI	Tecnologia da informação
UML	Linguagem unificada de modelagem – <i>Unified modeling language</i>
UTEC	Unidade técnico–científica

1. INTRODUÇÃO	12
1.1. PROBLEMA DE PESQUISA	16
1.2. OBJETIVOS	17
1.2.1. OBJETIVO GERAL	17
1.2.2. OBJETIVOS ESPECÍFICOS	17
1.3. JUSTIFICATIVAS	18
2. REFERENCIAL TEÓRICO	21
2.1. ONTOLOGIA	21
2.2. PROCESSAMENTO DE LINGUAGEM NATURAL	31
2.2.1. PRÉ-PROCESSAMENTO TEXTUAL	37
2.2.2. ANÁLISE LÉXICA	40
2.2.3. ANÁLISE SINTÁTICA	47
2.2.4. ANÁLISE SEMÂNTICA	57
2.2.5. ANÁLISE PRAGMÁTICA	71
2.3. RECUPERAÇÃO DE INFORMAÇÃO	74
2.4. INDEXAÇÃO AUTOMÁTICA	85
2.5. MARCO TEÓRICO	94
3. METODOLOGIA	97
3.1. TIPO DE PESQUISA	97
3.2. CARACTERIZAÇÃO DA AMOSTRA	99
3.3. INSTRUMENTO	106
3.4. PROCEDIMENTO	112
4. PROPOSTA DE ARQUITETURA	117
4.1. DESCRIÇÃO DA ARQUITETURA	117
4.2. AVALIAÇÃO DA ARQUITETURA	126
5. CONCLUSÕES	138
REFERÊNCIAS	144
APÊNDICE A – SCRIPT PYTHON	159
APÊNDICE B – ONTOLOGIA	164

Svenonius (2000) declara na abertura de sua obra que o acesso eletrônico instantâneo a informação em formato digital é o atributo mais distinto da era da informação. Assim como a máquina a vapor foi a entidade mais importante da era industrial, a representação do suporte tecnológico do período, a invenção do computador é a tecnologia que permite o sustentáculo da sociedade globalizada (CASTELLS, 2005), em rede, e da era da informação. Os elaborados mecanismos de recuperação de informação (RI) que são necessários para suportar o acesso à enorme quantidade de informação digital são produtos da tecnologia. A tecnologia sozinha, contudo, não é suficiente (SVENONIUS, 2000). Por si só, a tecnologia da informação (TI) não tem propriedades ou recursos suficientes para entregar resultados aceitáveis de RI em bases de dados não estruturadas, muito grandes e em constante crescimento, como as atuais.

Esse fato demanda, mais do que nunca, a associação de estudos de representação e organização da informação e do conhecimento com TI, a fim de extrair o que áreas de pesquisa, em princípio, distintas têm a se complementar para que a tecnologia não seja o fator que limite o desenvolvimento da informação. Esta pesquisa procura se inserir precisamente nesta discussão. Assim, o trabalho propõe uma arquitetura de indexação automática (IA) de documentos utilizando mecanismos de processamento de linguagem natural (PLN) em nível semântico. Nessa união que se alvitra entre a ciência da informação (CI) com a ciência da computação (CC), objetiva-se que os pressupostos de representação e organização da informação e do conhecimento da CI, neste caso particular aqueles voltados para indexação de documentos, sejam utilizados pela CC para automatização dos processos de tratamento da informação. Com isso consegue-se reduzir o alto custo que a indexação manual (IM) de documentos apresenta, mantendo-se a qualidade na RI que a indexação garante, encontrando um ponto de equilíbrio.

A questão subjacente trata da comoditização da TI. Ciências mais maduras, estabelecidas como área de pesquisa há mais tempo, como a engenharia elétrica, por exemplo, já estão em um nível de comoditização bastante avançado. Isso significa que os usuários de seus produtos quase nada têm de saber de seus

princípios para utilizá-los. Para fazer funcionar uma televisão, geladeira ou computador, por exemplo, basta liga-los a uma tomada. Produção de energia, transmissão de eletricidade, tensão, polaridade, e até a voltagem não são de conhecimento comum, e nem devem ser. A CC, entretanto, ainda é uma ciência muito jovem, e por isso ainda obriga os usuários de seus produtos a ter um conhecimento por vezes avançado de suas questões. Sommerville (2011) declara que a engenharia de *software* é uma ciência imatura, com apenas 44 anos de evolução, e que, portanto, ainda não tem seus pressupostos tão bem estabelecidos ou empiricamente testados em comparação a outras engenharias. Isso implica que seus produtos finais não apresentam o mesmo nível de comoditização e usabilidade de áreas mais amadurecidas.

Em um sistema operacional (SO), por exemplo, para inicializar um programa é necessário apontar um dispositivo (mouse) para o canto inferior esquerdo da área de trabalho e clicar em 4 (quatro) ícones diferentes. Em outro SO, o apontamento deve ser no canto superior esquerdo. Em um terceiro, em uma barra de ícones que aparece na lateral esquerda. Para trocar essa disposição, exige-se em média 12 (doze) cliques de mouse. Analisando-se friamente, um usuário de computador nem deveria saber que existe uma entidade chamada SO, ou como operá-la. O que ele deseja é ligar seu computador na tomada e dizer: “Quero ir ao cinema!”, e o computador listar, instantaneamente, todas as projeções, horários e preços das salas de exibição a um raio de 10 (dez) quilômetros de distância de sua posição atual. Ou então: “Quero mandar uma mensagem para alguém!”, e o computador já se preparar para o ditado do destinatário e do texto. Não obstante a existência de vários sistemas que já atinjam esse alinhado grau de interface com seus usuários, essa ainda é uma área de pesquisa que demanda evolução, padronização, cobertura e melhoria na qualidade dos resultados atuais.

Além disso, conhecimento de redes de computadores é normalmente requerido. Se determinados documentos estão armazenados em algum dos computadores da rede, ou em algum compartilhamento em qualquer das dezenas de servidores que existem é um problema latente de grandes organizações atualmente. Isso obriga os usuários a conhecerem algo que deveriam completamente ignorar que existe: a topologia de uma rede de computadores. O que se deseja é que

qualquer informação esteja, resalte-se, instantaneamente disponível quando se fizer necessária. Fora a questão da usabilidade propriamente dita, o usuário comum ainda tem de saber a diferença entre um aplicativo, um editor de texto, uma planilha eletrônica, um *software* antivírus, como operá-los e qual a aplicação melhor adaptada para cada um deles.

A CC sozinha, assim como, em princípio, todas as outras ciências, não consegue tratar todas essas questões. Isso justifica as associações com outras áreas do conhecimento. Para melhoria da usabilidade, por exemplo, pressupostos da arquitetura são comumente utilizados. Para RI, foco desta pesquisa, a CI desempenha papel importante. A linguística exporta conhecimento da linguagem para melhoria das interfaces de comunicação. A web semântica é um assunto que tem objetivos claros para progresso da RI na Internet. PLN, de maneira geral, tem aplicabilidade direta para solução de vários desses problemas.

Ladeira (2010) desenvolve uma pesquisa para análise da produção científica da comunidade acadêmica nacional na área de PLN. Descobre-se, entre outros, que a maior parte da publicação nacional de trabalhos científicos em PLN ocorre a partir do ano 2000, com a CC e a linguística ocupando 85% (oitenta e cinco por cento) da produção. A CI apresenta, portanto, uma participação modesta. Além disso, estudos de indexação diminuem a partir da década de 1980. Percebe-se, destarte, que conquanto o PLN seja uma área de pesquisa amadurecida em língua inglesa, em língua portuguesa ainda é incipiente. Os estudos nacionais em PLN estão muito localizados, com 12 (doze) pesquisadores responsáveis por mais de 20% (vinte por cento) da produção nacional, sendo nenhum deles da CI, centrados principalmente nos estados de São Paulo (SP), Rio de Janeiro (RJ) e Rio Grande do Sul (RS). Isso é uma lacuna muito grande, e esta pesquisa pretende contribuir para seu preenchimento.

O PLN tem potencial para encaminhar várias soluções para as questões levantadas. Em relação à usabilidade, a compreensão de linguagem falada é uma interessante área de pesquisa em PLN, com análise de áudio e reconhecimento fonético. Além dessa, as pesquisas de tradução automática (TA) têm potencial para universalizar o acesso a informação. Já em relação à RI, embora IA seja uma área

de pesquisa madura, acredita-se que a combinação do conhecimento profundo dos métodos linguísticos com o poder de extensão dos métodos estatísticos oferece uma arquitetura robusta. Ademais, o aprofundamento da análise linguística ao nível semântico representa o rompimento da barreira que há em pesquisas de IA, em vários idiomas, e de PLN, em língua portuguesa, cuja grande maioria fica em nível sintático. Esta investigação se ocupa dessas questões. Dentre as contribuições inéditas da pesquisa para a ciência da informação se encontram a arquitetura de IA proposta, o conjunto de ferramentas que se integrou para sustentar tal arquitetura, e a ontologia de aplicação construída para análise semântica de PLN em idioma português, no domínio de crimes cibernéticos.

Este documento de pesquisa organiza-se da seguinte forma: O Capítulo 1, de introdução, tem por alvo explicitar as motivações que implicaram o estudo. Além disso, formula-se a situação problema propondo uma questão de pesquisa. O objetivo geral é determinado, e os objetivos específicos são delineados de forma a antever os produtos que são gerados pela pesquisa. Por fim, justifica-se porque esta pesquisa faz parte do corpo de conhecimento da CI, e quais as contribuições que ela oferece para a área.

Já o Capítulo 2, de referencial teórico, focaliza o posicionamento desta pesquisa na literatura da área. O fim é fundamentar as decisões tomadas ao longo dos experimentos a partir de pesquisas existentes e pressupostos aceitos pela comunidade científica. O referencial teórico contempla os conhecimentos das áreas mais diretamente relacionadas a esta pesquisa, quais sejam ontologia, PLN, RI e IA.

O Capítulo 3 trata do estudo metodológico da pesquisa. Nele a mesma é devidamente classificada para formalização científica. A amostra não probabilística é detalhada para reconhecimento do tipo dos documentos utilizados, sua macroestrutura, e a estratégia utilizada para amostragem. O grande conjunto de instrumentos utilizados no trabalho é então descrito. Cada uma das ferramentas empregadas ou desenvolvidas é detalhada e os mecanismos de integração são explicitados para, assim, demonstrar a arquitetura proposta. Concluindo, os procedimentos executados são esquematizados com o objetivo de permitir que o experimento seja avaliado e replicável.

O Capítulo 4, de proposta de arquitetura, é onde o resultado da pesquisa é detalhado. A arquitetura de IA com análise semântica de PLN consiste em uma das contribuições inéditas deste trabalho, e sua descrição, aplicação, limitação e potencial são discutidos.

O Capítulo 5, por fim, trata da conclusão da pesquisa. Nele são demonstrados os cumprimentos dos objetivos propostos. Além disso, as limitações do estudo são discutidas e as novas investigações que surgem a partir dos resultados alcançados são propostas.

O documento conclui com a lista de todas as referências utilizadas no trabalho, além de apêndices contendo o código fonte produzido em linguagem de programação Python e a ontologia de aplicação construída.

1.1. PROBLEMA DE PESQUISA

O problema de pesquisa enquadra-se na representação e organização da informação e do conhecimento. A produção de documentos tanto na esfera governamental quanto na empresarial aumenta cada vez mais ao longo dos anos. Esse material não é organizado por meio de uma arquitetura de informação adequada, tornando muito difícil, senão impossível, sua precisa recuperação. Com o avanço da TI voltado para a digitalização de documentos, o caos existente nas prateleiras está sendo convertido para os servidores de banco de dados (BD) corporativos, os quais possuem restrita informação parametrizada para identificação de documentos.

Percebe-se que a IM melhora significativamente os índices de RI em pesquisas textuais, no entanto seu alto custo inviabiliza sua extensa utilização nesses ambientes desorganizados e em constante crescimento. Assim, a IA surge como proposta para manutenção da qualidade do processo de RI controlando seus custos. Nesse aspecto, o aprofundamento linguístico em nível semântico, por meio da utilização de PLN, pode aquilatar a qualidade dos descritores selecionados. Esta

pesquisa insere-se nessa área procurando propor uma solução para a representação e RI em organizações usando PLN para IA semântico-ontológica.

A questão de pesquisa, assim, é investigar se uma arquitetura de IA suportada por análise semântica de PLN tem potencial para oferecer índices considerados ótimos, na literatura da área, para resultados de consultas.

1.2. OBJETIVOS

Nesta seção explicitam-se o objetivo geral e os objetivos específicos desta pesquisa a fim de fixar o azimute para condução dos trabalhos.

1.2.1. OBJETIVO GERAL

Propor uma arquitetura para IA de documentos não estruturados em língua portuguesa, com abordagem estatística e linguística e PLN em nível semântico apoiado por ontologia.

1.2.2. OBJETIVOS ESPECÍFICOS

Desenvolver e integrar ferramental computacional para PLN do idioma português em nível semântico, com vistas à IA.

Construir uma ontologia de aplicação para organização e representação do conhecimento do domínio crimes cibernéticos.

Testar e avaliar a arquitetura proposta com base nos índices de precisão e revocação obtidos pela IA do *corpus* de pesquisa.

1.3. JUSTIFICATIVAS

O relacionamento do tema desta pesquisa com a CI, de maneira geral, e com a representação e organização da informação e do conhecimento, em particular, é discutido por vários autores. Borko (1968) analisa que a CI é uma disciplina que investiga os modos de processamento da informação para otimização do acesso e usabilidade. As pesquisas em RI, tais como esta, inserem-se nessa definição. Além disso, o mesmo autor continua colocando uma preocupação com o corpo de conhecimento relacionado, entre outros aspectos, à organização, armazenamento e RI. Por fim, o componente de ciência aplicada que a CI possui, o qual desenvolve serviços e produtos, é exatamente onde este estudo procura se inserir.

Já Le Coadic (2004) afirma que a CI, além de estudar as propriedades gerais da informação, se preocupa com a análise dos processos de uso da informação e a concepção dos produtos e sistemas que permitam a comunicação, armazenamento e uso. Considerando que a RI tem importância para o uso da informação, esse trabalho se alinha a essa preocupação. Borko (1968) ainda ressalta a importância da investigação da representação da informação em sistemas artificiais. A indexação como processo de representação de documentos para fins de criação de pontos de entrada para a recuperação é discutida por Lancaster (2004). Robredo (2005) ressalta a importância da utilização de um modelo de representação do conhecimento embasado em tesouro para melhoria dos resultados da indexação. A pesquisa de Camara Junior (2007) prevê que a utilização de um esquema de representação do conhecimento mais elaborado, como uma ontologia, pode melhorar ainda mais a qualidade da indexação.

Brookes (1980), por sua vez, discute que o conhecimento objetivo do mundo 3 de Popper é uma exposição filosófica das atividades práticas que definem os cientistas da informação. Além disso, ele ressalta que a acessibilidade do conhecimento objetivo é um problema não considerado por Popper, e que demanda da CI estudos teóricos e práticos para aprimoramento dos resultados. A compreensão do mundo 3, que é composto por material objetivamente produzido pelo homem comparando-se aos outros 2 mundos, precisa de maior estudo e é

grande a oportunidade que a CI tem para assumir tal responsabilidade. Esta pesquisa insere-se precisamente nessa discussão.

Uma pesquisa de IA com base em modelos cognitivos para obtenção dos resultados deve possuir um caráter analítico e avaliativo. Para isso, ela vai se basear em estudos comparativos, qualidade da indexação, modelagem, métodos de aplicação de testes e medidas de performance. Borko (1968) coloca essa como uma das nove categorias de projetos de pesquisa em CI. Além dessa, esta pesquisa igualmente se insere na categoria de análise da linguagem, que tem componentes tais como linguística computacional (LC), lexicografia, PLN, psicolinguística e análise semântica. Destarte, os estudos necessários para construção e utilização da ontologia que apoia o processamento semântico desta pesquisa fazem parte da área de pesquisa em CI. Guarino (1998) discute que a pesquisa na área de ontologia está sendo reconhecida em áreas tais como a organização e RI. A pesquisa de Souza (2006) é um exemplo desse tipo de abordagem, quando reconhece como hipótese que uma das principais frentes de atuação para tratamento da informação se preocupa com a exploração de informações semânticas e semióticas intrínsecas aos dados, conquanto sua pesquisa não utilize uma ontologia, porém um tesouro, para consecução dos resultados.

Saracevic (1995) argumenta em seu artigo que a CI é uma ciência interdisciplinar por natureza, cujas relações com várias disciplinas estão em constante evolução. Ele menciona que a RI é uma área cujos problemas são objeto dos maiores esforços e desprendimento de recursos em CI. Esta pesquisa trata exatamente disso, quanto a sua proposta de melhoria da RI. Acrescenta, ainda, a relação com a CC na produção de ferramentas, serviços e redes. Os algoritmos relacionados à informação, foco central da CC, e a natureza da informação, objeto de estudo da CI, são abordagens complementares. Por fim, recupera um ponto importante desta pesquisa, qual seja a utilização de inteligência artificial na ciência cognitiva. Evidentemente uma pesquisa em CI não objetiva exclusivamente a produção do ferramental tecnológico para determinado resultado, todavia o uso desse ferramental no estudo e uso da informação.

Além desses, Holland (2008) integra ao centro das pesquisas em CI disciplinas distintas tais como biblioteconomia, TI, sociologia, comunicação, CC e inteligência artificial. Esta pesquisa tem um forte caráter tecnológico, até por causa de sua proposta de automatização, a qual é área de pesquisa da CI quando objetiva não apenas os produtos computacionais construídos, mas também os novos fluxos informacionais gerados e a avaliação do impacto da nova metodologia na forma como a informação é recuperada e utilizada (acessada) pelo usuário final.

Concluindo, o relacionamento desta pesquisa com a organização da informação vai ao encontro do conceito cunhado por Wurman (2005) em 1975 de arquitetura da informação. O autor defende que arquiteto de informação é o agente que dá clareza ao que é complexo, tornando a informação compreensível para outros seres humanos. Já Rosenfeld e Morville (2002) definem a disciplina como o *design* estrutural de um espaço informacional construído para facilitar a execução de tarefas e o acesso intuitivo ao conteúdo. Além disso, eles também mencionam que arquitetura da informação pode ser definida como a arte e ciência de estruturar e classificar web sites e intranets para ajudar pessoas a encontrar e gerenciar a informação. Portanto, uma pesquisa que favoreça a RI apoia os resultados para melhoria do acesso, e encontra-se na área de pesquisa da representação e organização da informação e do conhecimento.

2. REFERENCIAL TEÓRICO

O referencial teórico tem por alvo discutir e analisar o embasamento teórico para a pesquisa. Creswell (2009) afirma que a revisão da literatura tem vários objetivos. O primeiro é compartilhar com o leitor os resultados de outras pesquisas que sejam correlatas ou próximas a esta. Já o segundo trata de relacionar este estudo ao diálogo mais amplo que haja na literatura da área, preenchendo lacunas ou estendendo resultados anteriores. O terceiro, por fim, procura o estabelecimento de um *framework* por meio do qual se defina a importância deste estudo, assim como um *benchmark* para comparação de resultados prévios. Os autores citados, destarte, possuem pesquisas relevantes na área de ontologia, PLN, RI e IA, os quais foram importantes para fundamentação das propostas.

2.1. ONTOLOGIA

Guarino (1998) sugere a distinção entre Ontologia, com 'O' maiúsculo, e ontologias, no plural e com 'o' minúsculo. A primeira é uma disciplina acadêmica da filosofia, que trata do estudo do ser enquanto ser, com suas propriedades e categorias. Guarino, Oberle e Staab (2009) acrescentam que é o ramo da filosofia que trata da natureza e da estrutura da realidade, independente de quaisquer considerações posteriores, inclusive, de sua efetiva existência. Isso significa que uma Ontologia de curupiras, ou outra entidade fictícia qualquer, por exemplo, faz todo o sentido. Nirenburg e Raskin (2004) argumentam que o nome mais adequado para ela é metafísica.

Já a segunda, com uso prevalentemente maior na CI e na CC (GUARINO; OBERLE; STAAB, 2009), é reconhecida como esquema de representação do conhecimento, ou seja, o foco do estudo abandona as entidades propriamente ditas e passa a priorizar o conhecimento humano sobre elas. A existência, então, é avaliada sob uma régua bastante pragmática que determina que, para sistemas computacionais, existe aquilo que pode ser representado. Nesse aspecto, Nirenburg e Raskin (2004) alegam que o estudo das ontologias pertence ao campo de atuação da epistemologia.

Guarino (1998) também coloca que a pesquisa em ontologia está se tornando cada vez mais abrangente na comunidade de CC, em áreas como inteligência artificial, LC e BD. Ao descrever ontologia como um conjunto de axiomas lógicos desenhados para estabelecer o significado intensional de um vocabulário, as demandas de componentes elegantes de PLN são atendidas. Já Nirenburg e Raskin (2004) tratam ontologia como a modelagem semântica e o fluxo de informação entre as entidades que compõem a realidade. A modelagem ontológica, destarte, significa a utilização de um esquema conceitual em um formato explicitamente definido. Angele, Kifer e Lausen (2009) aventam ontologia como um modelo conceitual, o qual é uma descrição declarativa e abstrata da informação para um domínio de aplicação, além de formas para produzir inferências a partir dessa informação, o que é fundamental para as aplicações de PLN desta pesquisa.

Guarino, Oberle e Staab (2009) ainda especificam que uma ontologia computacional é um meio formal de modelar a estrutura de um sistema, ou seja, as entidades relevantes, nomeadas conceitos, e as relações que emergem da observação delas que sejam úteis para atingir os objetivos do sistema. A estrutura principal de uma ontologia, portanto, é uma hierarquia de conceitos: uma taxonomia. A essa hierarquia são evidenciadas relações e associações as quais, em um paradigma lógico, são modeladas por meio de predicados.

Gruber (1993), por sua vez, define ontologia, em uma citação muito utilizada em pesquisas na área, como a especificação explícita de uma conceitualização. Ele posta que a conceitualização é uma abstração da realidade, o corpo de representação formal do conhecimento de uma determinada área, com seus objetos, conceitos e as relações entre eles. Ou seja, a ontologia é o conjunto de termos de um vocabulário controlado, e suas relações, por meio do qual o conhecimento é representado. Ainda acrescenta que a conceitualização pode ser extensional ou intensional. A extensão é o conjunto daquilo a que uma entidade se aplica, ou seja, uma conceitualização extensional demanda a lista de todas as relações conceituais possíveis de um domínio. Como isso é inviável, ou no mínimo muito difícil, na maioria dos casos, uma maneira mais efetiva de especificar as

conceitualizações é formalizar uma linguagem e restringir as interpretações dela de forma intensional por meio de axiomas.

Os requisitos para tal linguagem, na opinião de Antoniou e van Harmelen (2009), são sintaxe e semântica bem definidas, suporte eficiente a raciocínio e inferências, poder e conveniência de expressividade. A lógica de primeira ordem pode ser aplicada nesse sentido e oferece o rigor que a ontologia precisa para ser utilizada por sistemas computacionais. A lógica descritiva (BAADER; HORROCKS; SATTLER, 2009) ou a lógica de quadros (ANGELE; KIFER; LAUSEN, 2009) também têm aplicação. Em resumo, a ontologia é um conjunto desses axiomas, ou seja, uma teoria lógica que captura um modelo intensional correspondente a certa conceitualização.

Guarino (1998) refina a definição propondo que ontologia seja uma teoria lógica que agrega um comprometimento ontológico a uma conceitualização. Ainda insiste que uma completa interpretação semântica deve ser possível sobre toda e qualquer declaração ontológica formal. E divide o estudo em tipos de ontologias, tais quais as modeladas na Figura 1.

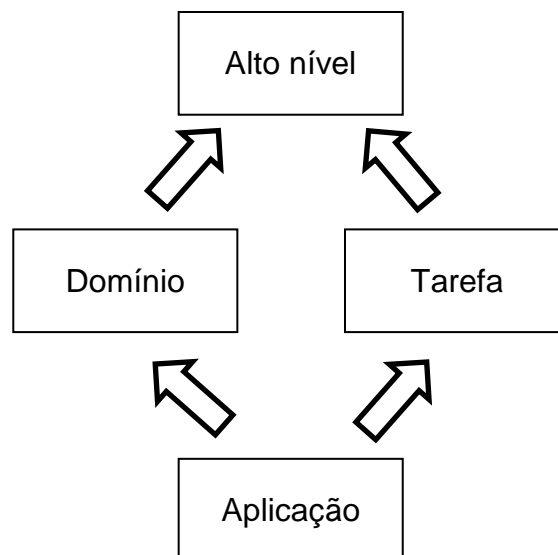


Figura 1 – Tipos de ontologias (GUARINO, 1998).

Uma ontologia de alto nível descreve conceitos gerais, independentes de um problema particular ou um domínio específico. A ontologia DOLCE (BORGIO; MASOLO, 2009) é um exemplo de ontologia de alto nível, com abrangência muito

grande, altamente reutilizável, bem fundamentada filosófica e conceitualmente, e semanticamente transparente, ou seja, ricamente axiomatizada. Uma ontologia de alto nível tem várias utilizações práticas, sobretudo para perspectivas de modelagem e engenharia de aplicações. Para PLN propriamente dito, entretanto, uma ontologia de alto nível não é aplicável, uma vez que o PLN está ligado a problemas específicos.

Já as ontologias de domínio, assim como as de tarefa, descrevem o vocabulário relacionado a elas por meio de especificação das descrições introduzidas na ontologia de alto nível. A diferença entre elas é que a ontologia de domínio focaliza a descrição das entidades do domínio, enquanto a ontologia de tarefa se preocupa com a descrição das atividades inerentes ao domínio. A ontologia para engenharia de *software*, proposta por Oberle, Grimm e Staab (2009) é um exemplo de ontologia de domínio. Os autores sugerem que a área é um bom candidato para uma ontologia de domínio, considerando que é suficientemente complexa, com diferentes paradigmas e aspectos, além de ser satisfatoriamente estável. Nela os autores descrevem os atores, artefatos, modelos, e demais objetos da área. Uma ontologia de tarefa nesse mesmo domínio observaria os processos envolvidos na engenharia de *software*, tais como levantamento de requisitos, modelagem, testes, entre outros.

Oberle, Grimm e Staab (2009) ainda acrescentam que a ontologia de domínio é apenas uma referência para o domínio. Para utilização computacional propriamente dita, essas ontologias devem ser reduzidas para ontologias de aplicação, em esquemas de representação do conhecimento, que são mais adequadas à operação. Essas ontologias de aplicação, então, descrevem conceitos que dependam de um domínio particular e de uma tarefa específica concomitantemente. São as ontologias mais especialistas e descrevem os papéis que as entidades do domínio assumem quando realizam determinada tarefa (GUARINO, 1998). Uma ontologia de aplicação é o tipo de ontologia que é construída e utilizada nesta pesquisa.

Uma vez que o conceito de ontologia encontra-se suficientemente claro, revelando-se um conceito ou definição operacional desta pesquisa, Sure, Staab e

Studer (2009) introduzem a disciplina de engenharia de ontologia, qual seja a que investiga os princípios, métodos e ferramentas para iniciar, desenvolver e manter as ontologias. O método proposto pelos autores deriva de diversos estudos de caso de construção e uso de ontologias na área de gestão do conhecimento. Pinto, Tempich e Staab (2009) agregam que um método deve preconizar, entre outros princípios, que as ontologias devem ser construídas de forma descentralizada. O argumento dos autores é que a centralização da construção de uma ontologia em uma única equipe perde detalhes do domínio que só são conhecidos e exercitados pelas áreas fins. Assim, as ontologias devem ser oferecidas à comunidade a que se destinam de forma que essa tenha certa autonomia sobre elas. Ainda acrescentam que as ontologias possuem um ciclo de vida que envolve interações entre a construção, modificação e uso. E, por fim, devem agregar a participação de não especialistas em seu processo de engenharia. Embora sejam reconhecidas as vantagens de desenvolver uma ontologia descentralizadamente, mormente quanto à captura de diversos pontos de vista dentro do mesmo domínio, ainda assim há que se considerar a dificuldade de integração desses resultados. Por vezes pode haver axiomas conflitantes ou contraditórios, o que pode embaraçar a coerência e a consistência da ontologia.

Uma interessante pesquisa de construção de ontologias foi desenvolvida por Belghiat e Bourahla (2012). Os autores propuseram uma geração automatizada de ontologias utilizando a linguagem ontológica para web (OWL) a partir de diagramas de classe da linguagem unificada de modelagem (UML). Para isso, construiu-se uma abordagem de regras de transformação utilizando grafos. A ontologia gerada é bastante simples, considerando a quantidade restrita de informação que há em um diagrama de classes. Ponderando, contudo, que esse é o formalismo padrão na engenharia de *software* para modelagem de sistemas, percebe-se grande massa crítica para melhoria de resultados.

Kasama, Zavaglia e Almeida (2010), por outro lado, constroem uma ontologia de domínio nas áreas de nanociência e nanotecnologia com método semiautomático. Essa ontologia objetiva a organização do conhecimento da área para aplicação em sistemas computacionais, de maneira geral, e repositório léxico para PLN, em particular. A importância desse trabalho para esta pesquisa encontra-

se na forma utilizada para levantamento das relações semânticas, a qual se baseou na abordagem de léxico gerativo de Pustejovsky (1991) que é detalhada na Seção 2.2.4 sobre análise semântica de PLN.

Já Vrandečić (2009) recobra um importante aspecto da engenharia de ontologias, qual seja a avaliação. Menciona que é uma área de pesquisa ainda emergente e levanta critérios de qualidade. A acurácia é a avaliação do quanto a ontologia representa corretamente aspectos do mundo real. A adaptabilidade mede o quanto a ontologia reage, sem alterações estruturais, a pequenas mudanças nos axiomas. Quanto à clareza, observa-se como a ontologia consegue comunicar efetivamente o significado dos termos definidos. Já a completude verifica o quanto o domínio de interesse foi devidamente coberto. A eficiência computacional mede a complexidade algorítmica dos módulos de raciocínio automático que processam a ontologia com sucesso. A concisão avalia se há redundâncias ou axiomas irrelevantes. A consistência acusa contradições. A acessibilidade organizacional, finalmente, afere quão fácil é o processo de instalação e distribuição da ontologia na organização. Todavia, tais critérios são de difícil medida e avaliar quão boa é uma ontologia não é tarefa trivial. Os métodos podem reconhecer problemas já documentados, como herança múltipla, por exemplo, de forma a permitir descobrir se uma ontologia não está bem construída, o que já é um grande passo no reconhecimento da qualidade.

Guarino e Welty (2009) discutem o método OntoClean. Esse é um mecanismo formal para análise de ontologias que valida a adequação ontológica e a consistência lógica das relações taxonômicas. Quatro noções básicas são descritas e utilizadas para avaliação das ontologias, quais sejam a essência, a identidade, a unidade e a dependência, recuperadas de estudos da filosofia. O objetivo é reduzir as relações redundantes ou desnecessárias melhorando as inferências automatizadas. Normalmente, as heranças múltiplas nas relações taxonômicas, se não eliminadas, são reduzidas ao mínimo essencial, o que facilita análises computacionais.

Investigando as aplicações de ontologias, Baader, Horrocks e Sattler (2009) argumentam que elas têm se tornado cada vez mais importantes nas áreas

de gestão do conhecimento, integração de informação, sistema de informação (SI) cooperativo, RI, comércio eletrônico e, com grande expoente mais recentemente, web semântica. Bruijn *et al.* (2009) acrescentam os *web services* semânticos, objetos descritos por ontologias para descoberta, combinação e execução automática de serviços. Esse é um promissor resultado para a web semântica.

Já Pinto, Tempich e Staab (2009) atestam o uso de ontologias, de maneira geral, para melhoria da qualidade da comunicação entre computadores, entre pessoas e computadores, e entre pessoas. Almeida e Bax (2003) realizam extenso levantamento sobre projetos utilizando ontologias nas mais diversas áreas, tais como gestão do conhecimento, comércio eletrônico, PLN, RI na web, e educação. Os projetos em PLN, de particular interesse desta pesquisa, mencionam-se Oncoterm, Gazelle, Mikrokosmos, e Pangloss para TA, e ambiente de desenvolvimento gramático multilíngue (KPML), Ontogeneration, Penman, e Techdoc para geração automática de textos em linguagem natural.

Stevens e Lord (2009) resumem as aplicações de ontologias, especificamente na área de biologia, por meio da Figura 2. Centralizam a descrição como raiz para qualquer utilização que se faça de uma ontologia. Percebe-se que, por seu caráter geral, tais aplicações podem ser estendidas para quaisquer domínios aos quais sejam aplicadas ontologias. Sombream-se as aplicações mais diretamente envolvidas nesta pesquisa.

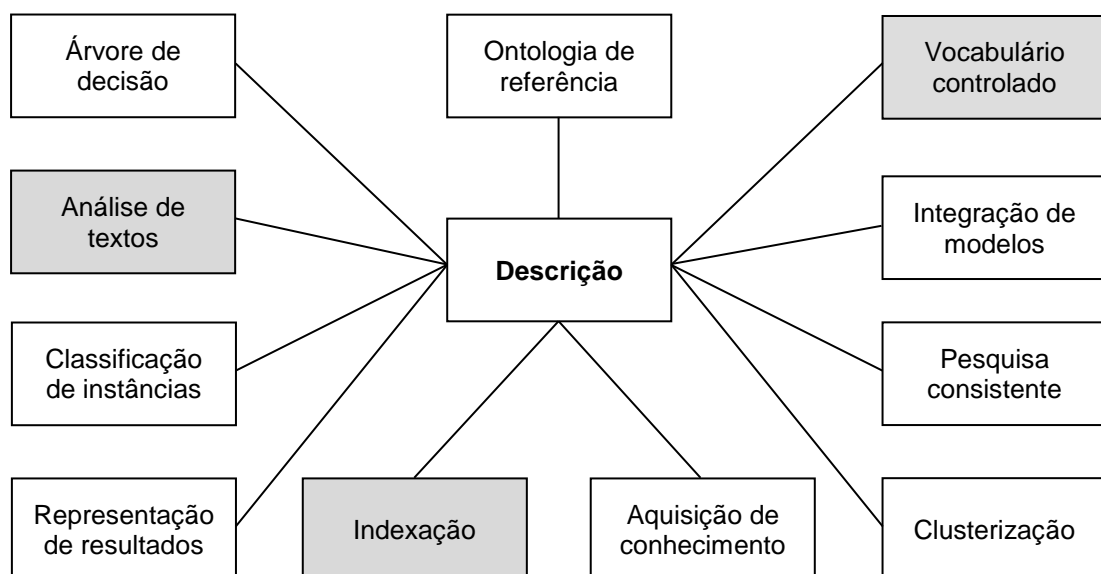


Figura 2 – Esquema de classificação para usos de ontologias (STEVENS; LORD, 2009).

Stevens e Lord (2009) detalham cada uma das aplicações propostas. Uma ontologia empregada como ontologia de referência define as classes de entidades de um domínio, o que tem utilidade por si só uma vez que a explicitação do conhecimento exige o questionamento de pressuposições implícitas no discurso da área. Além disso, serve de balizador para discussão, uma vez que é mais fácil argumentar sobre definições de entidades do que sobre palavras selecionadas como etiquetas das entidades. Já a utilização como vocabulário controlado, de grande importância para esta pesquisa, descreve as categorias de instâncias de conceitos utilizados para descrever o mundo. Com isso, cria-se um comprometimento na utilização de um vocabulário único para circunscrição de cada conceito, o que é muito útil em sistemas computacionais. Para integração de modelos, as ontologias podem ser aproveitadas não para comprometimento do vocabulário, como na utilização anteriormente descrita, mas no acordo das próprias categorias em que se dividem os conceitos do domínio.

Já a pesquisa consistente é uma aplicação de ontologias que se concerne na facilitação da pesquisa e análise da informação, como continuam Stevens e Lord (2009). Utilizando as propriedades da aplicação como vocabulário controlado, é possível restringir parâmetros de consulta para melhorar os resultados das pesquisas textuais, além de utilizar a estrutura taxonômica da ontologia para recuperação de instâncias de determinada classe, o que retorna suas respectivas subclasses. O uso de ontologias para clusterização fortalece os algoritmos para formação de conjuntos de documentos além da análise estritamente estatística de coocorrência, permitindo análise semântica. Já a aquisição de conhecimento representa a emprego de ontologias onde elas provêm protótipos para os atributos das instâncias de forma que formulários possam ser estabelecidos para sua aquisição e organização. Ademais, dados podem ser transformados para obedecerem às diretrizes da ontologia formando, assim, uma base de conhecimento por sobre a qual axiomas podem ser extraídos.

A aplicação de ontologias para indexação também tem grande importância para esta pesquisa. Utilizadas como um tesouro para avaliação dos descritores selecionados, as ontologias oferecem um maior conjunto de relações

entre conceitos. Para representação de resultados, uso que Stevens e Lord (2009) afirmam ser mais recente, as ontologias permitem a descrição de resultados preliminares antes de sua efetiva publicação, possibilitando inferências ou outras formas de organização dos dados. Já para classificação de instâncias, uma ontologia descreve as classes de instâncias em um domínio, provendo conhecimento organizado em um conjunto de fatos sobre elas para reconhecer um elemento do domínio como membro de uma classe particular.

Stevens e Lord (2009) concluem com duas últimas aplicações. A primeira é análise de textos, completamente aderente a esta pesquisa, que trata do reconhecimento de papéis linguísticos, aplicações de mineração de textos, associação de palavras a conceitos, classificação de tipos de palavras para distinção ontológica. Os autores ainda acrescentam que o papel das ontologias nessa área é muito maior do que o escopo de seu capítulo, e esta pesquisa, em particular, focaliza a análise semântica de PLN com suporte ontológico. O último uso é para o processo de tomada de decisão por meio de árvores de decisão, capturando conhecimento sobre um domínio e encapsulando restrições sobre classes e seus membros. Assim, uma ontologia pode oferecer a um sistema, ou mesmo ao usuário diretamente, fatos discriminadores para distinção entre classes de entidades.

A pesquisa de Santos (2006) demonstra a utilização das ontologias para integração de SI. Sistemas distintos, com BD modelados diferentemente, apresentam grande dificuldade para serem integrados. A construção de ontologias que permitam a interpretação dos dados de forma a possibilitar o tráfego de informações entre aplicações traz ganhos tais como economia de recursos, aumento da consistência, e melhoria da robustez por meio de replicação, entre diversos outros.

Já a pesquisa de Moreira (2010), em uma discussão das fronteiras entre a CI e a CC, procura investigar os subsídios teóricos e práticos que as ontologias têm a oferecer para a construção de informações documentárias em SI documentário. Conclui-se que esses não podem prescindir das tecnologias desenvolvidas para a engenharia de ontologias. Além disso, são indagadas as contribuições das ontologias para a construção de tesouros, e as contribuições dos tesouros na

construção de ontologias. Demonstrou-se que tesouros e ontologias são modelos que não possuem a mesma natureza, uma vez que têm concepções e finalidades distintas, conquanto ambos sejam esquemas de representação do conhecimento em contextos de linguagens especializadas.

Hage e Verheij (1999), por outro lado, propõem uma ontologia que integre todas as áreas do direito analisando-o como um sistema dinâmico e interconectado de estados das causas. O sistema é dinâmico porque as leis mudam, contratos são assinados, direitos são adquiridos. Ele é interconectado porque os elementos da lei não são independentes entre si, mas se conectam por meio de regras tais como “roubo demanda punição”, ou “assinatura de contrato leva a obrigações”. A pesquisa descreve um modelo por meio do qual a causa sofre a intervenção de um evento o qual altera seu estado anterior para um novo estado. Por meio desse tipo de abordagem é possível realizar inferências subliminares não explícitas no texto, ou seja, na descoberta de um estado e um evento, via PLN, pode-se inferir novo estado. Esse é um resultado promissor para o componente semântico.

Hirst (2009), por fim, propõe que as ontologias, tais quais objetos não linguísticos que representam mais diretamente o mundo, podem prover uma interpretação ou base para o sentido de palavras. Uma forma simples de fazer isso é produzir um mapeamento entre o sentido de unidades lexicais a elementos ou estruturas de uma ontologia. Isso só funciona, evidentemente, na extensão do quanto a ontologia consegue capturar a essência dos significados. O autor ainda menciona que aplicações na área de TA têm muito a usufruir dessas propriedades. Esse é um resultado fundamental para esta pesquisa, que é utilizado no módulo semântico de PLN.

Concluindo o referencial teórico de ontologias, o mesmo se associa a esta pesquisa em dois aspectos preponderantes. O primeiro refere-se ao tópico de engenharia de ontologias, uma vez que o escopo do trabalho determina a construção de uma ontologia de aplicação para organização e representação do conhecimento de um domínio. Os princípios e pressupostos da área são empregados para execução desse objetivo específico. O segundo aspecto é a aplicação de ontologias para representação e extração de conhecimento. A análise

semântica de PLN que se propõe para IA exige um esquema de representação do conhecimento que seja capaz de oferecer suporte a extração de significado. A ontologia construída atende a esse requisito.

2.2. PROCESSAMENTO DE LINGUAGEM NATURAL

PLN é uma área de pesquisa que durante muito tempo apresentou duas abordagens clássicas para discussão de seus problemas: as abordagens linguísticas e as abordagens estatísticas. Nos últimos 10 anos, contudo, o que se percebe é que a incorporação de conhecimento linguístico no processamento estatístico tem se tornado cada vez mais comum. Dale (2010) defende esse ponto de vista e acrescenta que a abordagem híbrida é provavelmente o futuro da pesquisa na área. Clark, Fox e Lappin (2010) concordam com ele acrescentando que ao contrário das abordagens estarem em conflito, os seus pontos fortes podem ser integrados para se complementarem. As técnicas linguísticas oferecem representações compactas de informação de alto nível, o que geralmente ilude os modelos estatísticos. Por outro lado, as abordagens estatísticas atingem um nível de robustez e cobertura que os métodos linguísticos raramente, senão nunca, conseguem sozinhos. Nesse capítulo, abordam-se pressupostos teóricos para ambos os modelos, de forma a permitir a construção das bases para suporte aos resultados.

Di Felippo e Dias-da-Silva (2009) constroem uma concepção de PLN segundo a qual a área é uma engenharia de conhecimento linguístico. Assim, exige-se uma sistematização dos métodos necessários para construção de sistemas de PLN, uma vez que há exigência de uma grande variedade de dados complexos necessários à simulação da competência e do desempenho linguísticos. O campo de pesquisa é privilegiado e fértil e, embora a área de investigação em língua inglesa seja mais madura que em língua portuguesa, grande potencial para crescimento há, mormente quanto à significativa evolução apresentada nos últimos anos. A Linguateca, por exemplo, é um repositório digital disponível *on-line* no sítio eletrônico de Internet em <http://www.linguateca.pt>. Trata-se de um centro de recursos para processamento computacional da língua portuguesa. Várias ferramentas estão disponíveis para utilização, além de publicações de projetos em

andamento e fórum de discussões para pesquisadores da área. A evolução da base de dados da Linguatca demonstra o quanto o fomento à pesquisa em PLN em língua portuguesa encontra-se em franca expansão, e esta pesquisa insere-se neste contexto.

Tradicionalmente, o trabalho em PLN tende a ver o processo de análise da linguagem como uma decomposição em estágios, tais quais as distinções teóricas da linguística, quais sejam a sintaxe, a semântica e a pragmática (DALE, 2010). A primeira trata da ordem e da estrutura. A segunda aborda o significado. Já a última reflete o significado contextualizado. A pragmática concerne com o discurso, enquanto as anteriores preocupam-se com questões sentenciais. Essa estratificação tem um propósito eminentemente pedagógico, uma vez que por vezes é bastante penoso separar o processamento da linguagem nas suas respectivas caixas. Cherpas (1992) concorda e afirma que não há consenso na extensão do quanto esses extratos devem ser separados, ou quais deles são primários e secundários. A estratificação constitui, entretanto, de base para modelos arquiteturais que tornam o PLN mais gerenciável do ponto de vista da engenharia de *software*.

Historicamente, porém, a granularização do modelo teve de ser refinada, produzindo uma decomposição mais detalhada para atender o tratamento de dados em linguagem real. A Figura 3 demonstra os estágios da abordagem linguística de análise em PLN, conforme apresentado em Dale (2010), iniciando o exame na superfície do texto, aumentando em cada passo a profundidade da análise.

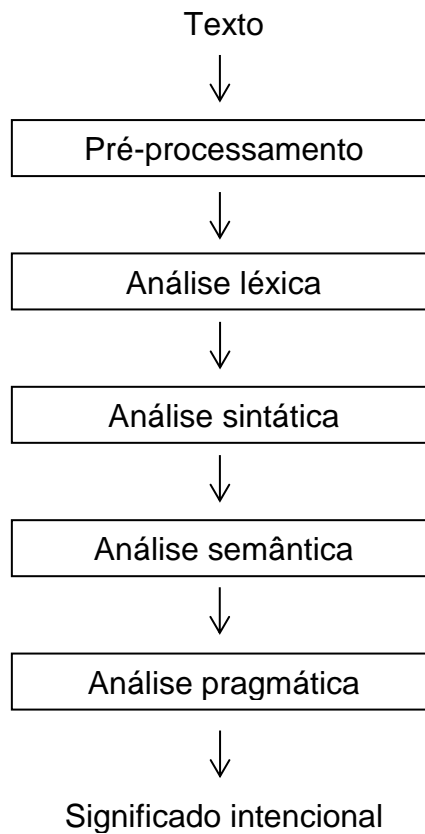


Figura 3 – Estágios de análise em PLN (DALE, 2010).

Nirenburg e Raskin (2004) concordam com essa estratificação em linha de produção, onde os resultados do passo anterior são precisamente as entradas do passo seguinte. Propõem, contudo, a possibilidade de que o conhecimento gerado por um módulo posterior seja retroalimentado a passos anteriores com o objetivo de facilitar a desambiguação. Cria-se, dessa forma, um ciclo de melhoria contínua a ser interrompido pelo atingimento de nível de qualidade proposto ou por uma regra de parada. Assim, desenham a Figura 4, onde as setas maiores representam a sequência dos passos para o processamento, e as setas finas representam a retroalimentação de conhecimento adquirido.

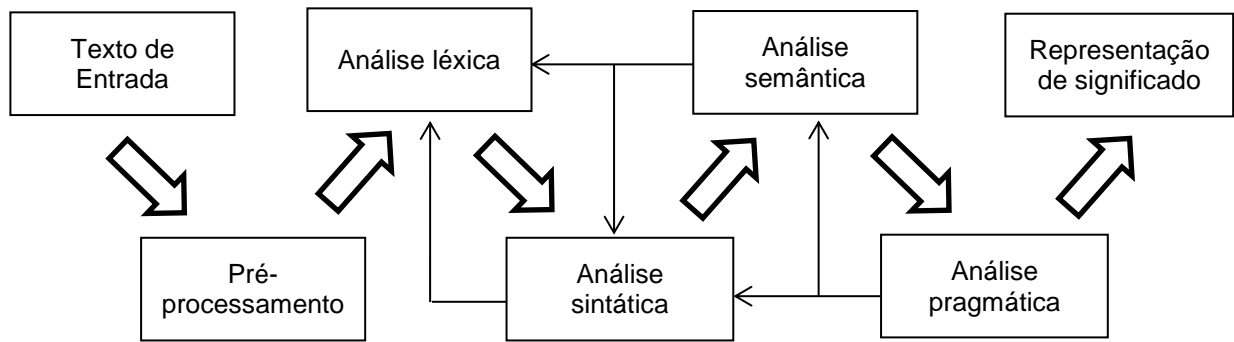


Figura 4 – Estágios retroalimentados de análise em PLN (NIRENBURG; RASKIN, 2004).

Há, também, uma abordagem de PLN dita plana, onde todos os módulos operam simultaneamente sem esperar pelo resultado do passo anterior. Nirenburg e Raskin (2004) sugerem a arquitetura modelada na Figura 5.

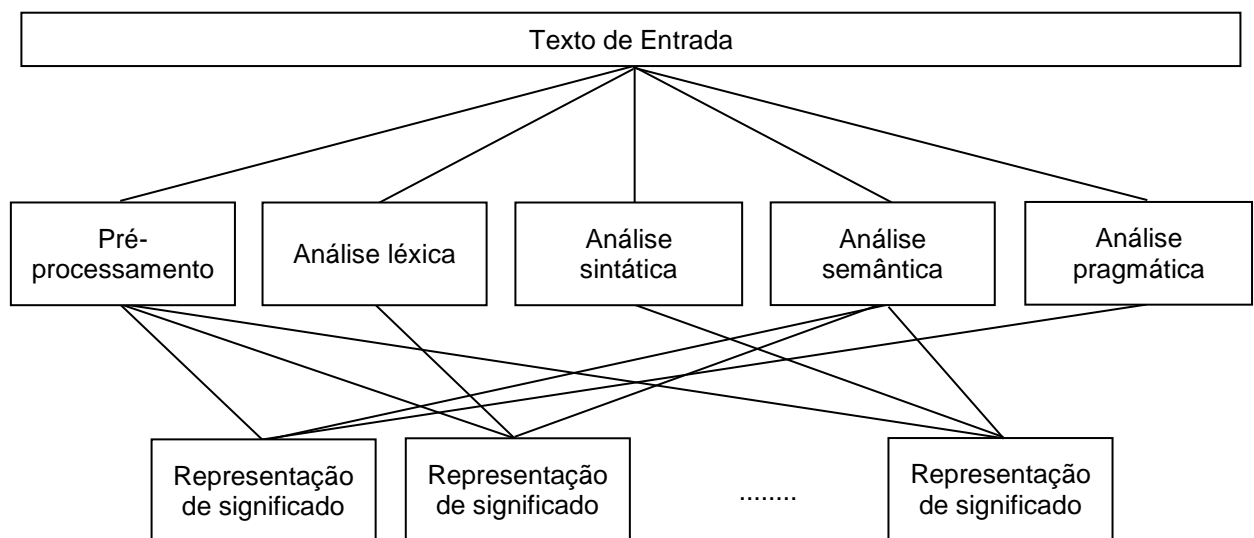


Figura 5 – Modelo plano de análise em PLN (NIRENBURG; RASKIN, 2004).

Note-se que para cada uma das representações de significado possíveis, não necessariamente todos os módulos de PLN são utilizados. É possível também separar o texto original em pedaços, e cada um desses ser tratado concomitantemente. Nesta pesquisa, vai-se optar por uma abordagem linear retroalimentada para PLN objetivando extrair as vantagens que ela intrinsecamente possui.

Percebe-se que a fase de pré-processamento, segmentação de unidades lexicais, e segmentação de sentenças, é o primeiro passo fundamental para início do

trabalho (CHAUDIRON, 2007). Jurafsky e Martin (2008) acrescentam que há um passo anterior, qual seja a fonética e fonologia, sobretudo em aplicações de tratamento de sons. Em sistemas que analisem estritamente documentos escritos, essa fase não existe ou é legada menor importância. Em linguagens como o português ou inglês, espera-se que o processo de reconhecimento de palavras seja facilitado pela separação de espaços em branco. Línguas orientais, entretanto, tais como o japonês ou tailandês, por exemplo, apresentam um complicador bastante maior para determinação do segmento. Palmer (2010) estabelece que a primeira tarefa seja precisamente a clara definição dos caracteres, palavras e orações do documento a ser analisado. Ainda acrescenta que o desafio se apresenta diferentemente em relação à linguagem que será analisada e a fonte do documento.

Na análise léxica, o objetivo é estudar a morfologia das unidades lexicais, ou palavras, e recuperar informação que será útil em níveis mais profundos de análise (HIPPISEY, 2010). A decomposição das palavras, assim como a detecção de regras de formação, permite a economia de espaço de armazenamento e aumenta a velocidade de processamento, considerando a hipótese simplista de armazenar cada unidade lexical encontrada em um repositório. Além disso, Dale (2010) ainda observa que sempre há a possibilidade de se deparar com uma nova palavra, não reconhecida pelo repositório. Nesse caso, o processamento morfológico pode oferecer estratégias para esse tratamento.

Já a análise sintática é aquela que se preocupa com a estrutura das sentenças em uma gramática formal. Um pressuposto em vários trabalhos de PLN é o de que o significado não se encontra nas palavras, mas sim na frase (LJUNGLÖF; WIRÉN, 2010). Kuramoto (1999) argumenta que o princípio básico do processo de indexação é que o sintagma nominal é melhor estrutura léxica que contém significado com qualidade para seleção de descritores, comparando-se a escolha de palavras isoladas, posição com a qual Chaudiron (2007) concorda. Souza (2006) e Maia (2008) extraíram sintagmas nominais em suas pesquisas para IA e classificação de documentos, respectivamente. Savoy e Gaussier (2010) também colocam que estruturas como adjetivo–substantivo, ou substantivo–substantivo, melhoram as dimensões da indexação. Como as orações não são apenas um amontoado de unidades lexicais, a análise gramatical e o *parser* sintático são

importantes para o reconhecimento. Essa área é provavelmente a mais estabelecida no campo de PLN (DALE, 2010).

A análise semântica trata do significado da sentença. Goddard e Schalley (2010) apresentam que, partindo da organização realizada pela fase anterior, qual seja a análise sintática, um objeto mais estruturado para manipulação e extração é gerado, permitindo a perscrutação do entendimento. Não é possível dar significado ao conteúdo, mas é possível analisar as relações válidas entre as palavras, a partir de seus conceitos, como bem coloca Chaudiron (2007). Essa análise demanda um forte esquema de representação do conhecimento para ser efetivada. Uma ontologia que determine traços semânticos das unidades lexicais é demandada para reconhecimento da validade das relações. Além disso, Nirenburg e Raskin (2004) declaram que é necessário um formalismo para representação do significado textual e conhecimento estático para processamento semântico, o qual inclua o mapeamento das estruturas de dependências sintáticas e semânticas, tratamento de referências e regras da estruturação textual.

O componente pragmático, por fim, procura incluir o contexto à análise linguística, a fim de permitir a geração de um significado. Esse utiliza uma base de dados construída em um esquema de representação de conhecimento para representar o contexto externo do texto e permitir a utilização desse conhecimento para inferências automatizadas (MELLISH; PAN, 2008). Mitkov (2010) defende que o discurso deve ser analisado como um todo, uma vez que a declaração não se concentra exclusivamente em uma oração, mas sim no conjunto delas. Nirenburg e Raskin (2004) listam que um módulo de análise pragmática não pode prescindir de um analisador de dependências em nível de discurso, um gerenciador de conhecimento contextual para interpretação, um interpretador de objetivos e planos do autor, e um determinante de estilo do texto. Pesquisas com esse grau de complexidade são muito escassas, a ainda não se apresentam em língua portuguesa.

Resnik e Lin (2010) ventilam um importante aspecto de PLN, qual seja a avaliação de resultados. Eles afirmam que o PLN é o braço de engenharia da LC e, nesse aspecto, se concerne da criação de artefatos computacionais para realização

de tarefas. A questão operacional na avaliação de algoritmos ou sistemas de PLN, destarte, é a extensão na qual os mesmos produzem os resultados para os quais foram construídos. Como os aplicativos de PLN raramente são construídos com uma arquitetura monolítica, ou seja, são formados por vários módulos diferentes, cada qual com objetivos particulares, usualmente alinhados em linha de produção, dificilmente uma única medida ou forma de avaliação pode ser bem aplicada. Ainda assim, é fundamental a preocupação com o conjunto de medidas de eficiência, tanto automática para apoio ao processo de construção, quanto contando com a participação de usuários finais, para medida final de efetividade.

Concluindo, a complexidade no aprofundamento da análise aumenta consideravelmente. Dale (2010) argumenta que os resultados alcançados atualmente em nível léxico/morfológico e sintático são bastante mais significativos do que em nível semântico e pragmático/discurso. Prossegue-se, então, com o estudo detalhado de cada fase do PLN com suas respectivas peculiaridades, desafios e estado da arte de soluções propostas.

2.2.1. PRÉ-PROCESSAMENTO TEXTUAL

A fase de pré-processamento do texto é aquela onde um arquivo de texto cru, normalmente apresentado como uma sequência de bits digitais é reconhecido como uma sequência bem definida de unidades linguisticamente significativas. No nível mais baixo, caracteres representando os grafemas do sistema de escrita da linguagem. Evoluindo, unidades lexicais consistindo de um ou mais caracteres. Por fim, sentenças contendo uma ou mais palavras (PALMER, 2010). O pré-processamento é fundamental para o PLN uma vez que essas unidades são utilizadas por todas as fases subsequentes do processo.

O pré-processamento pode ser dividido, de maneira geral, em dois processos. O primeiro é a triagem documental, que é a conversão de um conjunto de arquivos digitais em um documento de texto bem formado. O formato original dos documentos e a codificação dos caracteres são relevantes para esse passo. Além disso, é fundamental o reconhecimento da macroestrutura textual para descarte de

elementos não desejados, como imagens, cabeçalhos, marcação HTML, dentre outros. Palmer (2010) explica que a conclusão desse estágio é um *corpus* bem definido, pronto para ser utilizado para análises mais profundas. Xiao (2010) ainda acrescenta que o *corpus* precisa ter um tamanho adequado para a necessidade de treinamento ou aplicação para o qual se destina, frente à quantidade de documentos disponíveis. Além disso, a amostragem deve ser estatística e representativa de todo o conteúdo, para não criar vieses no ferramental.

O segundo processo, por outro lado, é a segmentação do texto. Esse estágio consiste na conversão do *corpus* em suas unidades lexicais e sentenças. A segmentação das palavras demanda o reconhecimento dos limites das mesmas. Em língua portuguesa, o delimitador evidente é o espaço em branco, porém outros caracteres especiais também podem ser usados para isso. Cada unidade lexical é nomeada *token*, e esse processo denomina-se tokenização (CHAUDIRON, 2007). Goldsmith (2010) detalha diversos métodos para segmentação de texto em *tokens*, inclusive para linguagens onde o espaço em branco não é o delimitador clássico, ou naquelas onde, em alguns casos, não existe o delimitador. Lista, ainda, várias abordagens para aprendizado morfológico não supervisionado, ou seja, sem a necessidade de *corpus* anotado para treinamento prévio do ferramental.

Já a normalização do texto é exatamente o processo de canonizar diferentes *tokens* para o seu lema original, com o objetivo de diminuir o espaço de armazenamento e acelerar o processamento. Bird, Klein e Loper (2009), por exemplo, explicam como realizar essa tarefa usando o ferramental de linguagem natural (NLTK) desenvolvido em linguagem de programação Python para qualquer idioma com base de treinamento disponível. O NLTK é um conjunto de ferramentas criado em 2001 como parte de um curso de LC do Departamento de CC e Informação da Universidade da Pensilvânia. Desde então ele tem sido suportado e expandido, servindo de base para vários projetos. Erjavec e Dzeroski (2004), por sua vez, desenvolveram um método estatístico para lematização de unidades lexicais em idioma esloveno. Para palavras conhecidas, o *framework* retornou 98.6% (noventa e oito ponto seis por cento) de acurácia. Para palavras não conhecidas, o ferramental produziu 92% (noventa e dois por cento) de efetividade, o que é um resultado bastante significativo.

Após o reconhecimento das palavras, passa-se ao reconhecimento das orações. Para isso, é necessário reconhecer o limite das frases, o que em português normalmente ocorre com o ponto final. As idiossincrasias tais como a utilização do ponto final para determinar a parte fracionária de um número, ou uma abreviação, por exemplo, devem ser tratadas nessa fase. Além disso, uma frase com um único ponto final ao término pode possuir várias orações, separadas por vírgulas ou semivírgulas, e cada uma dessas sentenças apresentar uma ideia própria. Isso é algo que o sistema de PLN deve considerar, como alerta Dale (2010). Bird, Klein e Loper (2009) mostram que o NLTK possui ferramentas automatizadas para segmentação de frases que sopesam esses dificultadores.

Essa segmentação possui como regra geral a utilização do conjunto “ponto final – espaço em branco – letra maiúscula”. Várias abordagens são utilizadas com o objetivo de desambiguar a ocorrência do ponto final aumentando a quantidade das unidades lexicais analisadas antes e após o ponto. Aquelas que utilizam algoritmos treináveis aproveitam uma extração probabilística da base de documentos para treinamento e sintonização do ferramental, enquanto outra parte é empregada para avaliação do sistema. Estratégias estatísticas suportam a segmentação da base, permitindo a extensão de resultados. Palmer e Hearst (1997) desenvolveram um sistema com esses pressupostos que atingiu a expressiva taxa de 99% (noventa e nove por cento) de acerto com documentos não formatados do *Wall Street Journal*.

Palmer (2010) identifica alguns desafios do pré-processamento textual. O primeiro deles trata do tipo do sistema de escrita da linguagem. Há línguas iconográficas, como o japonês, nas quais um grande conjunto de símbolos individuais representa as palavras. Nas línguas silábicas os símbolos individuais simulam sílabas. Já as línguas alfabéticas, como o português, usam os símbolos para representação do som. A maioria das linguagens usa modelos silábicos ou alfabéticos. Na prática, entretanto, qualquer linguagem, em algum momento, pode utilizar os três sistemas. O símbolo \$, em língua portuguesa, ou então o %, dentre vários outros, contém um significado intrínseco: iconográfico. Isso é um problema que deve ser tratado na tokenização.

Outro problema é a questão da dependência. Um sistema de PLN depende do conjunto de caracteres utilizado na codificação. Tradicionalmente o padrão *Unicode* implementado no sistema UTF-8 é amplamente utilizado, e contempla um conjunto enorme de símbolos para atender qualquer sistema de escrita. Ainda assim, documentos em diferentes codificações podem inviabilizar o PLN se não tratado. Além dessa, a dependência da linguagem propriamente dita é evidente. Um sistema de PLN para língua portuguesa tem menor chance de sucesso para língua inglesa, e menor ainda para chinês, por exemplo, considerando as diferenças estruturais de cada sistema de escrita. Ainda há a dependência do *corpus*. Um sistema de PLN treinado para documentos de certo formato não apresenta bons resultados em textos livres, com seus erros e quebras gramaticais. Mesmo textos bem formatados, contudo com macroestrutura diferente, também não são bem interpretados (PALMER, 2010). Esses desafios, dentre vários outros, são o foco de ataque para construção de sistemas de PLN flexíveis e genéricos.

Percebe-se, para concluir, que do pré-processamento depende o sucesso do sistema de PLN como um todo. Com o aprofundamento da análise, os erros cometidos anteriormente se potencializam e impedem o progresso do processamento. Assim, é importante tratar todas as questões e conseguir boa taxa de acerto na tokenização e separação de sentenças. Mais do que um pré-processamento, essa fase deve estar fortemente integrada no *design* e implementação dos outros estágios do sistema.

2.2.2. ANÁLISE LÉXICA

A análise léxica, ou morfológica, se ocupa do estudo das palavras. Na fase anterior elas já foram devidamente tokenizadas e reconhecidas, permitindo-se aprofundar a análise sobre elas. Hippiisley (2010) atribui às palavras o status de tijolos para construção de textos em linguagem natural. Acrescenta ainda que há normalmente duas abordagens em PLN para tratamento de uma unidade lexical. A primeira é tratar simplesmente como uma *string* de texto: uma cadeia de caracteres. A segunda é considerar a palavra como um objeto mais abstrato, que é um termo

derivado de um lema canônico com um conjunto de regras de formação. Uma tarefa básica, nesse caso, é relacionar uma variação morfológica de uma unidade lexical ao seu lema original, o qual se encontra em um dicionário de lemas com suas respectivas informações sintáticas e semânticas. Essa abordagem é mais elegante e econômica, por conseguinte evidentemente mais utilizada. Heinecke *et al.* (2008) usaram um modelo como esse para seu sistema de PLN.

Percebe-se, portanto, que há dois processos. Chaudiron (2007) descreve o primeiro como o *parse* da unidade lexical, onde uma palavra encontrada é canonizada em seu lema, e armazenada no dicionário com as respectivas regras de formação. O segundo, denominado geração, é exatamente a volta do processo, ou seja, a partir do lema e do conjunto de regras chegar à palavra derivada. Para IA essa volta não é tão relevante, todavia para sistemas de TA, por exemplo, ambos os sentidos são fundamentais. Hippiisley (2010) ainda acrescenta que para sistemas de RI não indexados, o processo de geração é computacionalmente mais econômico do que listar parâmetros de pesquisa. E ainda melhora a revocação, considerando a existência de palavras novas que não sejam reconhecidas pelo dicionário, mas possam ser geradas por um analisador morfológico eficaz. Logo, é muito importante que o mecanismo utilizado para análise léxica seja flexível o suficiente para realizar os dois processos.

A noção ideal de que uma unidade lexical nada mais seja do que a soma de seu lema a sufixos que designem a derivação chega a muitos resultados, porém não contempla a riqueza da linguagem. Erjavec e Dzeroski (2004) argumentam que a língua eslovena, assim como diversas outras incluindo o idioma português, tem uma grande riqueza de inflexões, com substantivos flexionando para gênero e número e uma configuração complexa de terminações e modificações do lema original. Um exemplo bastante simples deixa esse conceito suficientemente claro. A regra geral, em língua portuguesa, para o plural é o acréscimo do caractere 's' ao final. O plural de 'ciência' é 'ciências'. Ocorre, contudo, que há vários plurais irregulares: o plural de 'informação' é 'informações', onde além do acréscimo do 's' há alteração de vogal no lema. Isso ocorre para praticamente todas as derivações em língua portuguesa. Hippiisley (2010) comenta que esse modelo é bastante completo para o finlandês, no entanto deixa a desejar para o inglês. Savoy e

Gaussier (2010) listam várias variações de unidades léxicas para diversos idiomas demonstrando que isso não pode ser descrito por meio de uma regra geral trivial.

Morfologistas reconhecem três abordagens clássicas para estruturação de palavras. A primeira é denominada item e arranjo. Ela atende o caso ideal onde uma unidade lexical é a derivação de seu lema somado a um sufixo. A segunda chama-se item e processo. Nela, é levado em consideração o processo por meio do qual palavras complexas são geradas pela variação do sufixo, por tipo de lema. A ênfase focaliza o processo fonológico que está associado à operação morfológica. A última abordagem, por fim, é chamada palavra e paradigma. Nela, o lema é endereçado em uma tabela que associa a variante morfológica do lema com o conjunto das propriedades morfossintáticas. Essa tabela normalmente é implementada como uma árvore de derivação (HIPPISEY, 2010).

Na construção de ferramentas de PLN para as duas primeiras abordagens, sendo que estudos há também para aplicações na terceira, o modelo computacional mais efetivo é a utilização de autômatos finitos, ou também chamados autômatos de estado finito (AEF), na opinião de Hippiisley (2010). Hopcroft, Motwani e Ullman (2006) explicam que AEF nada mais são do que um formalismo matemático por meio do qual se constrói uma máquina de estados que lê entradas de uma fita e transita entre eles. Para definir um AEF, é necessário lançar mão do conceito de tupla, qual seja uma sequência ordenada de um número limitado de objetos que se constituem para formar a estrutura de uma definição matemática. Assim, a definição formal de um AEF é precisamente a tupla descrita abaixo:

$$A = [Q, \Sigma, \delta, q_0, F]$$

Onde:

A : AEF.

Q : conjunto finito de estados possíveis.

Σ : conjunto finito de símbolos de entrada, também denominado alfabeto.

δ : função de transição que recebe o estado atual (membro de Q), um símbolo de entrada, que é parte do alfabeto (membro de Σ), e retorna o novo estado (membro de Q).

q_0 : o estado inicial do AEF (membro de Q).

F : conjunto finito de estados finais, também denominados estados de aceitação, do AEF (subconjunto de Q).

O estudo dos AEF é de fundamental importância para a CC. O reconhecimento de expressões regulares ou o desenvolvimento de compiladores usam basicamente esse formalismo. Por sua proximidade com a linguagem, considerando o alfabeto da fita de entrada, os AEF também são muito importantes para a LC. Para a CI, qualquer pesquisa textual bem implementada é suportada por um AEF, o que demonstra sua enorme utilidade prática. Wintner (2010) coloca que os AEF são dispositivos computacionais que geram linguagens regulares, no entanto eles também podem ser vistos como reconhecedores. Dado um AEF que gere uma linguagem, e uma palavra qualquer, é fácil determinar se tal palavra pertence àquela linguagem, em tempo linear. Isso justifica o uso de AEF para uma aplicação simples de PLN, porém imprescindível, qual seja a procura em dicionários, que só é computacionalmente possível usando esse formalismo. A forma mais tradicional para representação de um AEF é um grafo direcionado, todavia Jurafsky e Martin (2008) mostram que uma tabela de transição de estados também resolve o problema menos elegantemente, desperdiçando espaço de armazenamento, mas com eficácia.

Na análise léxica, em particular, os transdutores de estado finito (TEF) são muito utilizados tanto para a abordagem item e arranjo quanto para a item e processo (HIPPISEY, 2010). Esses transdutores são autômatos não determinísticos, aqueles onde a função de transição não pode estabelecer um único estado de destino após determinado símbolo de entrada. A diferença consiste em que no TEF há duas fitas de símbolos: uma para entrada e outra para a saída. Vários transdutores podem ser compostos para permitir a formalização de qualquer tipo de derivação de lemas de uma estrutura gramatical. Jurafsky e Martin (2008) argumentam que os TEF têm uma função mais geral que um AEF uma vez que esses definem uma linguagem formal por meio do reconhecimento de cadeias de caracteres, enquanto aqueles também modelam as relações entre conjuntos de cadeias. Ainda resumem as funções de um TEF como reconhecedor, gerador, tradutor ou relacionador.

Wintner (2010) mostra, por exemplo, a utilização de um TEF para definição de singulares e plurais de unidades lexicais, mesmo naquelas onde há grande variação do lema, usando as duas fitas de símbolos. Goldsmith (2010) também discute a utilização de TEF para análise léxica. Adiciona, contudo, um que contém mais de duas fitas, para ser utilizado em linguagens com morfologia mais complexa, concordando que o TEF tradicional de duas fitas seja plenamente completo para o inglês. Porter (1980) propõe um algoritmo bastante simples, entretanto extremamente eficiente e, portanto, ainda o mais utilizado atualmente para lematização. Esse algoritmo usa um TEF para modelar regras de reescrita não demandando treinamento supervisionado prévio, demonstrando, com isso, seu grande valor.

Uma importante ferramenta da análise léxica é o etiquetador *part-of-speech* (POS). A abordagem consiste em partir de uma sentença completa, com seu conjunto de palavras, e etiquetar cada unidade lexical encontrada com sua respectiva categoria morfológica ou classe. Güngör (2010) declara que esse é um subprocesso da análise morfológica, ou processo complementar, porquanto a análise léxica propriamente dita envolve encontrar a estrutura interna de uma palavra, seu lema, suas derivações e regras de formação, enquanto o etiquetador POS faz uma análise superficial, associando uma categorização à unidade lexical. Um exemplo é o etiquetador morfológico usado na pesquisa de Camara Junior (2007) ou Heinecke *et al.* (2008) que tinha por objetivo definir se uma determinada unidade lexical, dentro de uma determinada oração, era um substantivo, adjetivo, verbo, advérbio, etc. Bird, Klein e Loper (2009) prescrevem a utilização do etiquetador POS do NLTK nativamente para língua inglesa, ou utilização de *corpus* anotados para treinamento e utilização em outros idiomas.

Várias dificuldades se apresentam no processo de etiquetagem POS. O primeiro deles é a ambiguidade, onde palavras iguais podem assumir diferentes funções morfológicas dentro da sentença (JURAFSKY; MARTIN, 2008). Güngör (2010) dá exemplos dessa e ainda acrescenta o problema do encontro de palavras desconhecidas. O etiquetador deve ser capaz de utilizar a posição contextualizada da unidade lexical na sentença para prever sua classe. Ainda assim, a etiquetagem

é uma área muito pesquisada, e as ferramentas atuais apontam acurácia de 96% (noventa e seis por cento) a 97% (noventa e sete por cento), o que é um número significativo, principalmente frente às porcentagens de acerto das análises mais profundas, notadamente menores.

Em relação às abordagens, elas se dividem nas embasadas em regras e nas estatísticas, como concordam Jurafsky e Martin (2008) e Güngör (2010). Os primeiros ainda discutem abordagens híbridas, onde às regras é somado um componente estatístico de aprendizagem computacional. Nos etiquetadores embasados em códigos um conjunto de regras de transformação de etiquetas existe e vai sendo recursivamente aplicado aos dados até que a margem de erro abaixe ao nível proposto. Esse conjunto aumenta de acordo com o aprendizado de um *corpus* anotado. Os resultados experimentais são bastante satisfatórios, porém exige-se aprofundado conhecimento linguístico para compilação manual das regras.

As abordagens estatísticas surgiram exatamente para tratar essa desvantagem. Nelas, modelos estocásticos são construídos para aprendizado sobre grandes *corpora* anotados. Assim, a etiquetagem ocorre por probabilidade e os resultados experimentais também atendem às expectativas. A grande vantagem dos modelos estatísticos está na portabilidade, bastando o treinamento do *framework* em outros *corpora* para obtenção de resultado semelhante. Güngör (2010), Nivre (2010) e Jurafsky e Martin (2008) detalham o modelo de Markov e o modelo oculto de Markov (MOM). Aqueles ainda mencionam a máquina de suporte vetorial, as redes neurais, algoritmos genéticos, árvores de decisão, TEF, e vários outros mais utilizados atualmente em pesquisas de PLN em diversos idiomas.

Daelemans e Bosch (2010) discutem os modelos de aprendizagem embasados em memória. Essa é uma abordagem inspirada em trabalhos anteriores a Chomsky (1956) nas áreas de categorização psicológica e reconhecimento de padrões. Ela preconiza que a generalização pode também ser alcançada sem a formulação de representações abstratas, como as regras de uma gramática livre de contexto (GLC), a qual é detalhadamente discutida na Seção 2.2.3 de análise sintática. Schmid (2010), por outro lado, arrazoa o uso de árvores de decisão para etiquetagem em PLN. Justifica seu uso pelas vantagens na velocidade de

treinamento e processamento sobre grandes *corpora*. Além disso, a existência de um maduro conjunto de ferramentas computacionais torna fácil a aplicação.

Henderson (2010), por sua vez, particulariza o uso de redes neurais para aplicação em etiquetagem POS. Para modelagem estatística, a arquitetura de rede neural mais adequada é o *perceptron* multicamadas. Ela surgiu como resposta aos críticos do algoritmo do *perceptron*, os quais argumentavam que ele só podia ser utilizado em uma classe muito limitada de problemas. O algoritmo aprende a discriminar classes de saída a partir de combinações lineares das propriedades de entrada. Isso se encaixa muito bem na arquitetura de uma rede neural, e com a especificação de uma rede síncrona simples atingiu resultados de 90.1% (noventa ponto um por cento) de precisão em uma base de testes do *Wall Street Journal*.

Já Malouf (2010) elucida as abordagens de máxima entropia com suas múltiplas aplicações nas áreas de detecção de final de sentenças, etiquetagem POS, resolução de ambiguidade em árvore de derivação, TA, entre outras. Bird, Klein e Loper (2009) relatam brevemente o MOM e citam o modelo de campos aleatórios de cadeia linear condicional. Jurafsky e Martin (2008) detalham como esse MOM é uma abordagem de máxima entropia para classificação de sequências. Oferecem ainda a formulação estatística, aplicações e demonstração do quanto esse mecanismo é importante para PLN. A observação de quanto esse modelo é citado e utilizado em pesquisas da área corrobora a afirmação. Wintner (2010), por fim, explica como utilizar um TEF para formalização de uma relação regular entre as unidades lexicais de duas linguagens: a linguagem natural e a linguagem das etiquetas POS. Ou seja, para cada unidade lexical da linguagem natural realizar a associação regular com a respectiva etiqueta POS dentro do contexto analisado.

Um exemplo de aplicação de etiquetagem POS em língua portuguesa é a pesquisa de Ribeiro, Oliveira e Trancoso (2003). Os autores desenvolvem um etiquetador morfossintático para o português europeu com resolução de ambiguidade léxica. A pesquisa é interessante no aspecto de utilizar abordagem híbrida: são usadas regras linguísticas para estruturação do etiquetador e uma formulação probabilística é empregada no motor estatístico para desambiguar as classificações morfológicas. O *corpus* é dividido em base de treinamento e teste, e

os resultados experimentais chegam a expressiva taxa de 97.07% (noventa e sete ponto zero sete por cento) de acerto para reconhecimento de substantivos e 96.93% (noventa e seis ponto noventa e três por cento) para verbos.

Já Garcia e Gamalho (2010) adaptam uma suíte de aplicativos denominada FreeLing que contém módulos de tokenização, segmentação de orações, e etiquetagem POS. O analisador morfológico utiliza um dicionário de língua portuguesa europeia e galega. Foi desenvolvido utilizando o MOM e um *corpus* anotado revisado por linguistas, o que apesar de seu tamanho reduzido apresenta alta qualidade na anotação morfossintática. Assim, os resultados de pesquisa apresentaram valores de precisão próximos do estado da arte, justificando a importância dos *corpora* de treino para qualquer ferramental. Embora esta pesquisa seja em língua portuguesa europeia, o mesmo *framework* pode ser aplicado ao português brasileiro substituindo-se o dicionário morfológico e o banco de anotações sintáticas.

Concluindo, a análise léxica/morfológica permite aprofundar a análise de PLN e fornecer insumo à análise sintática, fase posterior. A utilização de TEF resolve grande parte dos problemas de linguagens com morfologia mais pobre, como o inglês, e composições mais complexas são modeladas para bem atender morfologias mais ricas, como a língua portuguesa. Em relação à etiquetagem POS, conquanto os resultados já possuam uma taxa de acerto bastante alta, por se tratar de uma tarefa basilar, pequenas melhorias têm potencial para produzir grandes alterações na qualidade de análises mais profundas.

2.2.3. ANÁLISE SINTÁTICA

A análise sintática é aquela onde uma sequência de unidades lexicais, tipicamente uma oração, será decomposta para determinar sua descrição estrutural de acordo com uma gramática formal (LJUNGLÖF; WIRÉN, 2010). O *parse* sintático normalmente não é um fim por si só, todavia um meio para a seleção de descritores (CAMARA JUNIOR, 2007) ou extração de significado, por exemplo. Para isso, utilizam-se os resultados das fases anteriores, quais sejam as unidades lexicais

tokenizadas e classificadas. O resultado da análise sintática é uma hierarquia sintaticamente estruturada preparada para interpretação semântica.

Primeiramente, é instrutivo especificar a diferença entre um processo de análise sintática de uma linguagem de programação computacional e uma linguagem natural. Uma linguagem de programação possui uma GLC bastante simples e rígida, sempre impeditiva de ambiguidades. Além disso, a complexidade computacional para o *parser* é linear ao tamanho da entrada. Por fim, uma GLC para linguagens de programação é completa, ou seja, uma sentença correta sempre pode ser decomposta, por definição.

Já nas linguagens naturais, a gramática é contextualizada, o que transforma o problema em complexidade exponencial. O problema da ambiguidade existe e demanda tratamento, pois a distribuição das possibilidades aumenta o espaço amostral. Nesse caso, inferências estatísticas podem oferecer algum tratamento para a contenção (NIVRE, 2010). Por fim, uma linguagem livre possui ilimitadas possibilidades de construção, inclusive incorretas, porém inteligíveis, o que dificulta a percepção se um determinado erro encontrado na análise é decorrente de um erro de construção ou falta de cobertura da gramática.

Chomsky (1956) define GLC como um conjunto de símbolos e regras de derivação entre eles. A definição formal segue na tupla abaixo:

$$G = [\Sigma, N, S, R]$$

Onde:

G : GLC.

Σ : conjunto finito de símbolos terminais da gramática.

N : conjunto finito de símbolos não terminais da gramática.

S : o símbolo inicial da sentença (membro de N).

R : conjunto de regras de produção.

Define-se também o conjunto $V = N \cup \Sigma$, ou seja, o conjunto de todos os símbolos reconhecidos pela gramática. Nota-se, portanto, que a regra de produção

R, também denominada regra de derivação (WINTNER, 2010), é escrita como $A \rightarrow \alpha$, onde $A \in N$ e $\alpha \in V$. Essa definição é tão simples e elegante quanto poderosa, uma vez que se tornou o mais influente formalismo para descrição de sintaxe de linguagens. Chomsky (1969) define também que gramáticas gerativas nada mais são do que um sistema de regras que, de forma explícita e bem definida, atribui descrições estruturais a sentenças. Jurafsky e Martin (2008) delineiam várias dessas regras de gramática formal para a língua inglesa permitindo um paralelo para outras linguagens.

Outra definição importante é a forma normal de Chomsky (FNC). Uma gramática é dita que se encontra normalizada quando cada uma de suas regras de produção obedece a uma das seguintes proposições: ou a regra é unária do tipo $A \rightarrow \alpha$, onde $A \in N$ e $\alpha \in \Sigma$, ou é binária do tipo $A \rightarrow B C$, onde A, B e $C \in N$. A normalização de uma gramática é sempre possível para reconhecimento de uma mesma linguagem, ou seja, se há uma GLC que reconhece uma linguagem, é sempre possível se descrever outra GLC que obedeça a FNC e reconheça a mesma linguagem. Hopcroft, Motwani e Ullman (2006) demonstram, entretanto, que isso altera radicalmente a gramática, e aumenta exponencialmente a quantidade de regras de produção. Wintner (2010) ainda expõe que há algumas flexibilizações da FNC, que não chegam a quebrar o paradigma, mas são úteis para alguns formalismos.

Wintner (2010) também explica que a forma padrão de representação da estrutura sintática de uma sentença gramatical é uma árvore sintática, árvore de derivação ou árvore de *parse*. A escolha de uma árvore é evidente, uma vez que as regras de produção derivam símbolos não terminais até os terminais, em quantos passos forem necessários para atingir o objetivo final, qual seja que todas as folhas da árvore sejam apenas símbolos terminais. Ljunglöf e Wirén (2010) acrescentam que conquanto GLC sejam completas para modelagem de linguagens de programação, linguagens naturais demandam formalismos mais complexos. Clark (2010) explica que há outros, tais como a gramática de generalização frase-estrutura, gramática léxico-funcional, gramática de junção-arbórea, gramática categorial combinatória, dentre várias. Demonstra-se por indução finita, contudo, que

todas elas são equivalentes a GLC ou alguma extensão de GLC, o que é um resultado importantíssimo para a LC.

Nederhof e Satta (2010) discutem algumas dessas gramáticas úteis para desambiguação das possibilidades sintáticas de uma oração. A primeira é a gramática léxico-funcional, que incorpora um elemento léxico aos símbolos não terminais da GLC. Esse elemento tem um papel importante no conteúdo sintático e semântico da *string* derivada. Outra é a gramática de junção-arbórea a qual é tão poderosa que atinge certos graus de sensibilidade ao contexto, ou seja, ultrapassa os limites de uma GLC. A GLC síncrona, por fim, é muito utilizada em aplicações de TA entre linguagens diferentes. O objetivo é sincronizar as regras de produção da linguagem de origem com as mesmas regras na linguagem de destino, probabilizando as ambiguidades. Assim, a derivação de uma GLC síncrona é um par de árvores de *parse* amarradas.

Acrescente-se uma breve menção à gramática de cláusula definida. Essa é um mecanismo nativo da linguagem de programação Prolog detalhadamente descrito por Blackburn e Bos (2005) que também é uma extensão de GLC. Essa gramática, em particular, por causa do poder da linguagem Prolog no sentido do aceite de múltiplos parâmetros, permite que se agreguem anotações semânticas aos itens léxicos de maneira muito direta. Além disso, todos os conectivos lógicos já se encontram implementados, tornando essa uma ferramenta muito útil para PLN, tanto em nível sintático quanto em nível semântico.

Uma vez que o conceito de GLC está suficientemente claro, parte-se para a definição formal de *parser*. Um *parser* é um decompositor, ou também nominado reconhecedor, que tem por objetivo analisar uma determinada sentença da linguagem e gerar uma árvore de derivação com as unidades lexicais tokenizadas da linguagem (CHAUDIRON, 2007). Nederhof e Satta (2010) definem *parser* como o processo de análise automática de uma sentença, sob a perspectiva de uma sequência de palavras, com o objetivo de determinar suas possíveis estruturas sintáticas. Para isso, é necessário um modelo matemático da sintaxe da linguagem, que são precisamente as GLC. Chomsky (1969) demonstra a construção da árvore,

discute as regras de produção e seus tipos, analisa a performance, transformações gramaticais e as fronteiras entre as análises sintática e semântica.

Um decompositor pode ser classificado como *top-down* se inicia sua análise a partir de símbolos não terminais até atingir os terminais nas folhas. Ele será *bottom-up* caso a análise se dê no sentido oposto, juntando os *tokens* das folhas para formar símbolos não terminais até a raiz. Jurafsky e Martin (2008) discutem vantagens e desvantagens de ambos os métodos. Outra dimensão de classificação é quanto ao determinismo de tratamento das ambiguidades. Caso uma decisão tenha de ser tomada para desambiguar uma questão, e apenas uma escolha irretorquível possa ser tomada, o *parser* é dito determinístico. Uma última dimensão trata da ordem de processamento do reconhecedor. Ele pode proceder da esquerda para a direita, ou seja, do início para o fim da oração, ou então pode trabalhar de dentro para fora, iniciando nos membros mais importantes dos sintagmas: o verbo no sintagma verbal e o principal substantivo no sintagma nominal (LJUNGLÖF; WIRÉN, 2010).

Alguns tipos de *parsers* são descritos por Bird, Klein e Loper (2009). O de descendência recursiva é aquele que adota uma abordagem *top-down*, recursivamente decompondo cada sentença até chegar aos símbolos terminais. O *shift-reduce* é um reconhecedor que utiliza uma pilha para alocação de cada símbolo terminal encontrado substituindo o topo da pilha pela composição dos símbolos. Percebe-se claramente uma abordagem *bottom-up*. Já o decompositor *left-corner* é uma evolução, com abordagem híbrida, de um *parser top-down* com filtragem *bottom-up*. Assim é possível impedir a ocorrência de recursão infinita.

Vários algoritmos existem para implementação de *parsers* sintáticos, mormente aqueles que bem atuam em GLC. O algoritmo Cocke–Kasami–Younger (CKY), por exemplo, é detalhadamente explicado por Ljunglöf e Wirén (2010) e Jurafsky e Martin (2008) exclusivamente para gramáticas na FNC. Jackson *et al.* (2003) o aplicam em sua pesquisa com sucesso. Nederhof e Satta (2010) apresentam o pseudocódigo do CKY e discutem sua aplicação, complexidade e revés, qual seja o fato de se exigir uma gramática na FNC. Esse é um importante

algoritmo padrão utilizado por grande parte das pesquisas aplicadas de PLN em língua inglesa em nível sintático.

Em evolução, outro modelo clássico é a abordagem de realizar o reconhecimento por dedução. Um *parser* dedutivo é um processo onde regras de inferência são usadas para derivar declarações sobre o estado gramático de sentenças a partir de outras declarações. Essas declarações são chamadas itens, e as derivações são axiomas. Nederhof e Satta (2010) discutem o algoritmo de Earley, o qual não exige a FNC, aceitando qualquer GLC arbitrária. Também detalham seu pseudocódigo demonstrando que o algoritmo de Earley é um sistema dedutivo. Jurafsky e Martin (2008) concordam e acrescentam que esse é um importante algoritmo de abordagem *top-down* que utiliza programação dinâmica.

Já o decompositor *left-right* (LR) procura, evoluindo frente aos anteriores, tratar várias regras gramaticais concorrentemente, juntando subpartes comuns. Essa foi uma estratégia criada exclusivamente para linguagens formais, depois estendida para linguagens naturais. O formalismo matemático utilizado é um autômato de pilha (HOPCROFT; MOTWANI; ULLMAN, 2006) chamado autômato LR, ou tabela LR.

A pesquisa de Alencar (2011) é um exemplo de desenvolvimento de *parser* para a língua portuguesa. O autor produz um sistema em linguagem de programação Python para ser integrado ao NLTK. Com ele, é possível etiquetar morfossintaticamente sentenças submetidas ao motor a partir de textos irrestritos. A abordagem utiliza *corpora* anotados em idioma português para treinamento assistido do *framework*. É prototipada uma GLC e alguns algoritmos clássicos de *parse* são aplicados para avaliação dos resultados de pesquisa. Conquanto os reconhecedores utilizados apresentem certo grau de imprecisão na análise de algumas sentenças, sobretudo naquelas onde há neologismos ou itens não dicionarizados nos *corpora*, ainda assim o projeto é uma importante iniciativa para o ainda muito restrito repertório de ferramentas de PLN em língua portuguesa.

Já Martins, Hasegawa e Nunes (2012) desenvolvem, em 2002, no âmbito do Núcleo Interinstitucional de Linguística Computacional (NILC), um decompositor de sentenças em idioma português. O NILC é um grupo de trabalho criado em 1993

para pesquisa e desenvolvimento na área de LC e PLN. O grupo inclui cientistas da Universidade de São Paulo, Universidade Federal de São Carlos e Universidade Estadual Paulista de Araraquara. Esse *parser*, denominado Curupira, utiliza um léxico morfossintaticamente anotado para suporte e treinamento da ferramenta. Uma gramática completa, com aproximadamente 600 (seiscentas) regras, é produzida para inicializar os algoritmos de *parse*. Essa gramática é precisamente a maior contribuição do Curupira para esta pesquisa, em particular, pois é a gramática utilizada na construção do módulo sintático de PLN.

O reconhecedor envolve várias questões complexas. A primeira delas é a questão da robustez. Um *parser* é dito robusto quando consegue chegar a algum resultado mesmo quando recebe uma entrada que não se conforma com o que se esperaria naturalmente. Uma oração sintaticamente errada, por exemplo, é negada pelo analisador, mesmo que tenha conteúdo relevante (CHOMSKY, 1969). Um reconhecedor robusto consegue realizar inferências ainda que o autômato não chegue a um estado consistente. A pesquisa de Heinecke *et al.* (2008) chegou ao expressivo resultado superior a 95% (noventa e cinco por cento) de precisão na análise sintática usando um *parser* para língua francesa capaz de realizar correções linguísticas.

Outro problema é a complexidade computacional. A complexidade de decomposição para GLC é linear para o tamanho da entrada. Para linguagem natural, entretanto, a complexidade cresce exponencialmente. O algoritmo CKY, por exemplo, oferece complexidade quadrada. As gramáticas de junção-arbórea apresentam complexidade à sexta potência. Pratt–Hartmann (2010) e Ljunglöf e Wirén (2010) recuperam a complexidade algorítmica para *parsers* dessas e de várias outras gramáticas. Aquele também discute a complexidade computacional de alguns modelos semânticos. Os últimos ainda argumentam que a complexidade teórica não necessariamente se reflete nas aplicações práticas, que mais se aproximam do caso médio.

A última questão que se deseja mencionar é a ambiguidade. Embora durante a análise sintática não haja informação completa para desambiguação, tal como restrições contextuais, ainda assim é possível e desejável que o *parser* realize

uma seleção no espaço amostral. Mesmo um decompositor não determinístico pode, pelo menos, diminuir as possibilidades de construção da árvore. Para isso, as metodologias estatísticas de PLN para reconhecimento sintático se baseiam em inferências matemáticas sobre amostras de linguagem natural. Podem ser utilizadas para diversos aspectos da análise sintática, como analisado por Nivre (2010), porém são primordialmente úteis para tratar o problema da ambiguidade. Nesse aspecto, elas complementam e estendem os *parsers* linguísticos.

Várias técnicas existem para tratamento estatístico em PLN. Zhang (2010) explica que a maioria delas vem da aprendizagem de máquina, que é a disciplina da inteligência artificial ocupada do aprendizado a partir de dados não estruturados. Isso significa extrair informação, descobrir padrões, prever informação que esteja faltando ou, mais holisticamente, construir modelos probabilísticos dos dados. Chelba (2010) descreve a formulação para o modelo probabilístico, qual seja a fórmula recursiva que calcula a probabilidade de uma sequência de unidades lexicais ser decomposta e etiquetada corretamente.

Dois tipos de aprendizagem há: a supervisionada e a não supervisionada. A primeira se ocupa da tarefa de prever informação não existente considerando informação previamente analisada. Métodos estatísticos são empregados para construir regras de predição a partir de dados anotados. Zhang (2010) detalha alguns deles, tais como a rede *bayesiana*, a máquina de suporte vetorial e a regressão logística. Os dados anotados são classicamente armazenados em uma estrutura denominada banco de árvores de *parse* (HAJICOVÁ *et al.*, 2010), ou seja, um BD estruturado com sentenças e suas respectivas árvores de derivação sintática. Esses bancos podem ser produzidos manualmente, por linguistas, ou automaticamente, por algoritmos computacionais, e possuem precisos métodos de pesquisa e interfaces de entrada para algoritmos estatísticos (JURAFSKY; MARTIN, 2008). Já a aprendizagem não supervisionada focaliza o agrupamento de dados em *clusters*. As principais técnicas estatísticas são modelos misturados e algoritmo de maximização de expectativa.

Clark e Lappin (2010) avaliam a acurácia versus o custo das abordagens supervisionadas e não supervisionadas. Conquanto as supervisionadas tenham

apresentado resultados na ordem de 88% (oitenta e oito por cento) a 91% (noventa e um por cento) de precisão, enquanto as não supervisionadas encontram-se na faixa de 75% (setenta e cinco por cento) a 79% (setenta e nove por cento), o custo de treinamento delas é muito alto, principalmente considerando que o custo das não supervisionadas é zero. Eles preveem inclusive que por isso é razoável esperar um maior foco no desenvolvimento desses sistemas no futuro do trabalho em PLN, uma vez que caso a acurácia e cobertura desses métodos melhorem eles vão se tornar alternativas cada vez mais atrativas aos métodos supervisionados.

No *parser* estatístico, as GLC e suas extensões são parcialmente substituídas por modelos estatísticos treinados no *corpus* de dados. Por meio da captura de tendências de distribuição deles, os modelos podem atribuir valores às possíveis análises de uma sentença, facilitando a desambiguação, além de diminuir as restrições gramaticais, o que favorece a robustez. Há também a grande vantagem de melhora na portabilidade de domínio ou até linguagem. Isso é possível uma vez que os modelos estatísticos são treinados a partir dos dados, ou seja, se houver mudanças significativas no *corpus*, o *parser* sempre pode ser novamente adestrado. Nivre (2010) cita essas vantagens e comenta o prejuízo, qual seja na forma de treinar o modelo estatístico. Sempre é necessário um grande volume com anotação correta (HAJICOVÁ *et al.*, 2010) para preparação do *framework*, o que ocupa grande parte do esforço, como exemplifica a pesquisa de Camara Junior (2007) durante a fase de treinamento do *parser* morfossintático. Liu e Curran (2006), em tentativa de diminuir esse prejuízo, desenvolvem um *corpus* a partir de conteúdo disponível na internet com dez bilhões de palavras em língua inglesa. A avaliação, que consiste no treinamento de um mesmo reconhecedor utilizando 3 (três) *corpora* diferentes, apresenta bom resultado de 95.1% (noventa e cinco ponto um por cento) de acurácia na correção ortográfica sensível ao contexto.

Há várias representações sintáticas na análise estatística. Uma delas é a estrutura constituinte, onde uma sentença é recursivamente decomposta em segmentos menores até chegar às unidades lexicais devidamente classificadas. Clark (2010) alega que essa abordagem se adapta muito bem às GLC em sua representação arbórea. Outra representação é a estrutura de dependência. Nela as palavras são relacionadas por meio de uma relação assimétrica binária, que é

chamada de dependência, como descrito por Bird, Klein e Loper (2009). Já Nederhof e Satta (2010) acrescentam que essa é uma relação gramatical. A pesquisa de Nivre e McDonald (2008), por exemplo, aplica a uma estrutura de dependência um modelo integrado de grafos e transições, gerando resultado consistentemente melhor do que a aplicação dos modelos separadamente. McDonald *et al.* (2005), por fim, desenvolvem um *parser* sintático também na representação de estrutura de dependência montando um grafo. O objetivo é rodar um algoritmo de conversão em árvore para desambiguação sintática. A última representação que merece menção é a gramática categorial combinatória, a qual conecta a análise sintática e semântica a inferências por meio de cálculo lógico proposicional (CLARK, 2010), ou quaisquer outras regras formais de combinação (JURAFSKY; MARTIN, 2008).

Conceitualmente, os modelos de reconhecedores estatísticos são compostos de duas partes. A primeira é o componente gerador, o qual tem por finalidade mapear cada uma das entradas nas suas respectivas classificações possíveis. O segundo é o avaliador. Esse módulo avalia segundo um critério numérico estatístico as classificações ambíguas e elege a melhor. Nivre (2010) descreve os módulos e ensina que ambos componentes demandam treinamento do *framework*. Esse exercício é supervisionado, como explicado por Zhang (2010), quando as sentenças da base de aprendizagem já se encontram corretamente classificadas, enquanto o não supervisionado utiliza texto livre para treino. Os resultados de treinamento supervisionado ainda são consideravelmente melhores, porém, como já colocado anteriormente, a produção de classificação para treinamento é um gargalo para grandes *corpora*.

Acrescentando, portanto, as análises estatísticas às análises estritamente linguísticas, Booth e Thompson (1973) cunharam a definição de GLC probabilística, ou estocástica. Essa nada mais é do que uma GLC tradicional onde para cada regra de produção é associada uma probabilidade. Esse número é utilizado para ranquear os possíveis resultados de classificação com o objetivo de selecionar os de maior pontuação. Nivre (2010) descreve criteriosamente os cálculos estatísticos e de distribuição de probabilidade para tomada de decisão. Nederhof e Satta (2010) mostram os pseudocódigos para implementação do algoritmo para encontro de *parse* mais provável e uma alteração do CKY para incorporação da probabilidade.

Jurafsky e Martin (2008) sugerem um método para extração automática das probabilidades a partir de um banco de árvores de *parse*. Também acrescentam que uma GLC probabilística pode ser lexicalizada por meio do acréscimo de um cabeçalho lexical para cada regra de derivação, de forma a condicionar a probabilidade da regra. O algoritmo CKY com algumas extensões também pode ser aplicado para essa evolução.

Para concluir, percebe-se uma tendência de que os métodos híbridos têm potencial para melhorar os resultados individuais conseguidos pelas abordagens linguísticas e estatísticas, como acreditam Dale (2010) e Clark, Fox e Lappin (2010). Lahtinen (2000) e Souza (2006) utilizaram esse tipo de estratégia em suas respectivas teses de doutorado com sucesso. A análise sintática é o aprofundamento onde a maior parte dos trabalhos de PLN chega. Em língua portuguesa, as pesquisas não fogem à regra. O objetivo desta, em particular, é romper essa barreira e acrescentar o componente semântico à infraestrutura de PLN.

2.2.4. ANÁLISE SEMÂNTICA

Goddard e Schalley (2010) afirmam que o objetivo final da análise semântica, tanto para pessoas quanto para sistemas de PLN, é entender o enunciado: não apenas ler o que está escrito, porém compreender a declaração. Mais pragmaticamente falando, eles listam aplicações na área de RI, extração de informação, criação automática de resumos, *data mining*, TA, tratamento de parâmetros de pesquisa de usuários, sistemas de representação do conhecimento, dentre outros. Esta pesquisa procura aplicar esse ferramental para RI e IA.

Na linguística, a análise semântica trata do significado das palavras, expressões, orações completas e declarações contextualizadas, essa última mais próxima da pragmática. Isso significa traduzir a expressão original em alguma forma de metalinguagem semântica ou sistema de representação (GODDARD; SCHALLEY, 2010). A lógica filosófica, tanto de primeira ordem quanto modais e não monotônicas, tem forte influência no tratamento, entretanto não é abrangente o

suficiente para compreensão de linguagem comum. Atualmente, o consenso que a linguística produziu nas áreas de fonologia, morfologia ou sintaxe, por exemplo, não é tão estabelecido na semântica. Uma das poucas questões que são universalmente reconhecidas é que o uso de linguagem comum envolve a integração de conhecimento linguístico, convenções culturais e conhecimento do mundo real.

Cruse (2011) delimita o estudo da análise semântica como parte de várias disciplinas acadêmicas. Ele reconhece que há uma grande sobreposição entre elas, contudo cada abordagem possui algum diferencial único. A filosofia, por exemplo, particularmente a filosofia da linguagem, focaliza estudos do tipo ‘como é possível que qualquer coisa signifique alguma coisa?’, ou então ‘que tipo de relação deve haver entre X e Y tal que X signifique Y?’. Já a psicologia, na psicolinguística, se preocupa em analisar como o significado é representado no cérebro, ou quais mecanismos há para codificação e decodificação de mensagens. A sociologia, por seu lado, especificamente a sociologia da linguagem, perscruta o papel da linguagem na sociedade e na criação e manutenção das relações sociais.

Já a neurologia enfoca como o processo de significação acontece em nível neuronal. Ao contrário da psicologia, que percebe esse processo de forma macro, a neurologia vai avaliar as conexões entre neurônios e sua influência. A semiótica vê a linguagem como um sistema de sinais entre vários outros, e pesquisa quais propriedades há para demonstrar sua proeminência. A estatística, de grande importância para PLN, usa os grandes *corpora* disponíveis para análises estocásticas sofisticadas com o objetivo de contribuir para o estudo da linguagem e da significação. Interesse particular há nos padrões de coocorrência de unidades lexicais nos textos em linguagem natural. A linguística, por fim, também fundamental para PLN, apresenta três aspectos chave. O primeiro são as intuições semânticas do falante nativo, as quais constituem de fonte primária de dados para análise. O segundo é a importância de relacionar o significado às várias análises superficiais da linguagem. A última é o respeito reconhecido não exclusivamente à linguagem propriamente dita, mas também a todas as suas variações (CRUSE, 2011).

Tradicionalmente, a análise semântica é dividida em semântica léxica e semântica composicional, também chamada combinatória ou gramatical. A primeira

procura estabelecer o significado das unidades lexicais ou de combinações fixas de palavras. Cruse (2011) declara que, sob essa ótica, os substantivos e adjetivos, ou seja, palavras com conteúdo, são mais importantes do que preposições ou artigos, por exemplo. As combinações fixas, ou expressões, são discutidas por Baldwin e Kim (2010) e Constant, Sigogne e Watrin (2012). A segunda é focada na compreensão das infinitas possíveis combinações de unidades léxicas em frases que obedecem as regras gramaticais. Goddard e Schalley (2010) argumentam, contudo, que apesar da divisão facilitar a pesquisa, atualmente está se reconhecendo que ambas as áreas interagem e se interpenetram de várias formas.

O problema mais proeminente da análise semântica em PLN é a resolução da ambiguidade (GODDARD; SCHALLEY, 2010). Do ponto de vista computacional, uma declaração está aberta a múltiplas interpretações porque algumas palavras possuem mais de um significado, o que é chamado ambiguidade léxica. Além disso, alguns operadores, como quantificadores, modais ou negativos, podem se aplicar a diferentes áreas do texto, com distâncias indefinidas, o que se denomina ambiguidade de escopo. Por fim, a ambiguidade referencial descreve a possibilidade de um pronome ou qualquer outra unidade referencial não estar evidentemente estabelecida (BIRD; KLEIN; LOPER, 2009). Em relação à ambiguidade léxica, particularmente, é usual a distinção entre a homonímia e a polissemia. A homonímia é o fenômeno de palavras diferentes, com sentidos obviamente dessemelhantes, se apresentarem na mesma forma, podendo essa forma ser sonora ou escrita. O verbo 'parar' conjugado no presente do indicativo, na terceira pessoa do singular, e a preposição 'para' exemplificam o conceito. Já a polissemia trata da mesma unidade lexical carregar significados distintos em contextos díspares. O exemplo clássico é a fruta 'manga' e a 'manga' da camisa. Ambos os fenômenos são problemas para PLN, no entanto a polissemia é maior uma vez que a sintaxe é normalmente a mesma, com diferenças mais sutis, tornando a análise bastante tendente ao erro. Cruse (2011) detalha a polissemia em seus subtipos e oferece diversos exemplos para elucidação do conceito e de seus tratamentos linguísticos.

Yarowsky (2010) atesta que a desambiguação do sentido é essencialmente um problema de classificação. Dada uma unidade lexical qualquer

em determinado contexto, e um conjunto de possíveis anotações semânticas para ela, selecionar qual é a mais adequada é o objetivo. As palavras próximas àquela que se está analisando, bem como a função sintática delas, oferecem poder de predição. As anotações semânticas podem estar inventariadas em dicionários de domínios ou hierarquia de conceitos, ou esquemas de representação do conhecimento mais complexos que recuperem também as relações entre os conceitos. Jurafsky e Martin (2008) contribuem afirmando que há várias aplicações práticas para desambiguação, notadamente RI, sistemas de pergunta e resposta, classificação textual e TA.

Os algoritmos para desambiguação de sentido embasados em aprendizagem a partir do *corpus* encontram-se no espectro entre os completamente supervisionados e completamente não supervisionados (YAROWSKY, 2010). Os supervisionados utilizam dados anotados para treinamento, o que demanda grande esforço de produção como descrevem Hajicová *et al.* (2010), enquanto os menos supervisionados realizam suas inferências a partir de textos livres. Ambos utilizam uma fonte secundária de conhecimento, tal qual um esquema de representação. Essencialmente, qualquer algoritmo genérico de classificação por aprendizagem de máquina serve para a abordagem supervisionada. As abordagens menos supervisionadas, no entanto, demandam estruturas mais robustas, tais como algoritmos de navegação em grafos, clusterização vetorial ou clusterização aglomerativa.

Jurafsky e Martin (2008) explicam que, para as abordagens supervisionadas, as redes bayesianas oferecem um modelo estatístico bastante efetivo. A estratégia mais utilizada, contudo, é o algoritmo de Lesk, e suas variações, o qual utiliza as unidades lexicais vizinhas para apoiar a decisão de desambiguação. Os esforços de pesquisa na área de desambiguação semântica se concentram em um ponto de grande interesse nesta pesquisa, qual seja o desenvolvimento de algoritmos que consigam melhor utilizar conhecimento ontológico disponível em esquemas de representação formal para iniciar algoritmos não supervisionados em textos não anotados, o que se chama *bootstrapping*.

Abandonando a discussão específica da ambiguidade, reconhece-se mais holisticamente que várias abordagens teóricas há para PLN em nível semântico. A primeira que se deseja mencionar é a abordagem lógica. Essa acredita que o significado de uma expressão se determina por meio do significado de suas partes, e pela forma como essas partes se combinam, qual seja o princípio da composicionalidade de Frege, como concordam Goddard e Schalley (2010), Bird, Klein e Loper (2009), Fox (2010) e Jurafsky e Martin (2008). Nitidamente não há uma lógica universal que cubra todos os aspectos de significado linguístico, ou relacionamentos linguísticos. A lógica de predicados, entretanto, com a qual é possível expressar propriedades de conjuntos de objetos, e suas relações, por meio de predicados, conectivos lógicos, quantificadores, conclusões e inferências, é bastante conhecida e utilizada. É interessante notar que não há análise do significado dos predicados, que em PLN se referem aos itens léxicos, nem estudos precisos de temporalidade, porém perscruta-se o entendimento global por meio de suas relações e inferências. Isso demonstra que a abordagem lógica é do tipo semântico composicional.

Blackburn e Bos (2005) discutem detalhadamente a abordagem lógica sob uma vertente bastante técnica, utilizando a linguagem de programação Prolog. Eles estabelecem que há três tarefas sistemáticas para tradução de sentenças em linguagem natural para predicados de lógica de primeira ordem. A primeira é especificar uma sintaxe minimamente razoável para o fragmento de linguagem natural de interesse. A segunda é especificar a representação semântica para os itens léxicos. A última é a tradução composicional, isto é, deve ser possível traduzir toda e qualquer expressão por meio de composições das traduções de suas partes. Observe-se que por parte se define a subestrutura gerada pela árvore de *parse* sintático. Evidentemente, todas as tarefas devem ser estruturadas de forma a naturalmente serem conduzidas a uma implementação computacional.

Jurafsky e Martin (2008) argumentam que a abordagem lógica é a base de significação para a web semântica. Revelam, ainda, que web semântica é uma tentativa de prover um mecanismo de especificação formal de significado para o conteúdo da Internet. Já Lacy (2005) afirma que a web semântica é a próxima geração da web, que vai suportar processamento automatizado da informação. Um

componente chave para o mecanismo envolve a criação de ontologias para os domínios de interesse, e a linguagem usada para representar o conhecimento, integrar e interpretar essas ontologias é a OWL, como concordam Antoniou e van Harmelen (2009) e Hitzler e Parsia (2009). Ela incorpora a lógica descritiva, com seus gráficos de redes de representação, *framework* de descrição de recursos (RDF) (PAN, 2009) (HERTEL; BROEKSTRA; STUCKENSCHMIDT, 2009), regras, e outros pressupostos de semântica da abordagem lógica para atingimento dos resultados.

Outra abordagem semântica composicional chama-se teoria de representação do discurso. Goddard e Schalley (2010) explicam que essa surgiu a partir do pressuposto de que o significado emerge ao analisar o conjunto de orações, e não uma sentença destacada do texto. A ideia básica é que a leitura de um texto cria uma representação mental, e a ela são somadas cada uma das novas frases. Formalmente, o enfoque requer dois componentes. O primeiro é uma definição formal da linguagem de representação, a qual define uma estrutura de representação do discurso. O segundo, um procedimento recursivo de construção para extensão dessas estruturas a partir da chegada de novas informações. A estrutura, portanto, contém um conjunto de referentes do discurso, que são os objetos, e condições aplicadas a eles, escritas em lógica de primeira ordem. A abordagem é muito útil para tratamento de orações com sujeito indefinido, pressuposições e anáforas. Fox (2010) adiciona que um algoritmo de construção sistematicamente monta a representação dos indivíduos descritos no discurso com suas respectivas propriedades e relações entre eles. Jackson *et al.* (2003) desenvolvem um algoritmo suportado por esse enfoque para solução de anáforas com impressionante resultado de 62% (sessenta e dois por cento) de resoluções corretas, 33% (trinta e três por cento) de não resolução e apenas 5% (cinco por cento) de resolução erradas. Bird, Klein e Loper (2009) ainda evidenciam a utilização de bibliotecas nativas do NLTK para materialização dessa abordagem.

A última abordagem semântica composicional discutida é a análise semântica orientada à sintaxe. Jurafsky e Martin (2008) explicam que essa parte do resultado da análise sintática, qual seja a árvore de derivação, e conecta às regras da GLC instruções que especificam como computar a representação semântica de uma construção a partir do significado de suas partes constituintes. O objetivo é

produzir um mapeamento regra a regra entre sintaxe e semântica. Assim, um cálculo de predicados e argumentos permite a aplicação de lógica de primeira ordem para representação de significado. Os autores ainda apresentam o algoritmo de Earley para *parser* sintático modificado para permitir a análise semântica nesses pressupostos.

Já a abordagem de léxico gerativo, desenvolvida por Pustejovsky (1991), é um tratamento voltado para a semântica léxica. Ele reconhece que unidades lexicais assumem diferentes sentidos, facetas, em contextos particulares, e acredita que o significado pode ser derivado por meio de uma rica representação em níveis. O sistema computacional que suporta o modelo apresenta quatro níveis, a saber:

- Estrutura de argumento: especifica o número e o tipo de argumentos lógicos e como eles são sintaticamente realizados;
- Estrutura de evento: definição do tipo de evento de uma unidade lexical. Alguns exemplos de tipos de eventos são: estado, processo e transição, sendo possível a construção de hierarquia de eventos;
- Estrutura de explanação: contendo quatro tipos:
 - Constitutivo: do que é feito um objeto;
 - Formal: o que um objeto é, o que o distingue em um domínio;
 - Télico: qual é o propósito ou função de um objeto;
 - Agentivo: como o objeto foi concebido, fatores envolvendo sua origem.
- Estrutura de herança léxica: identificação de como a unidade lexical é relacionada a outras unidades, além de suas contribuições para a organização global do léxico.

Assim, há um mapeamento semântico de cada símbolo terminal de uma gramática, e um conjunto de regras de inferências permite a análise contextualizada do conteúdo. Essa é uma abordagem bastante diferente daquelas embasadas nos pressupostos lógicos. O objetivo é conseguir uma detalhada decomposição léxica para a semântica linguística.

Outra abordagem semântica léxica é a metalinguagem semântica natural. Essa teoria cognitiva é embasada no estabelecimento empírico do que se chamam primos semânticos. Esses são significados simples e indefiníveis que parecem estar presentes em qualquer linguagem. A metalinguagem usa um subconjunto padronizado da linguagem natural com seus respectivos significados e propriedades sintáticas. Goddard e Schalley (2010) mostram uma tabela com os primos semânticos para a língua inglesa e comentam que há a mesma tabela para vários outros idiomas. Por extensão, alegam que os primos semânticos são conhecidos por serem uma espécie de intersecção de todas as linguagens. Para conceitos complexos, agregam-se os primos semânticos, simples por definição, em moléculas semânticas: conceitos maiores que podem ser explicados via arranjo de primos semânticos. Embora a abordagem seja provavelmente o melhor modelo teórico para semântica léxica, ela teve pouca aplicação prática em PLN.

A última abordagem que se deseja brevemente relatar, essa também semântica léxica, é a semântica com orientação a objetos (OO). Área muito nova da semântica linguística, ainda tem atuação bastante restrita, principalmente focada na representação do significado de verbos. Pustejovsky (1991) previu que ao se combinar a estrutura de explanação de um substantivo com a estrutura de argumento de um verbo uma composição emerge tal qual um modelo de programação com OO. Considerando o grande corpo de pesquisa em OO, tanto para programação como para modelagem, essa é uma abordagem que tem potencial para crescer. Schalley (2004) introduz um *framework* para representação de semântica verbal embasado na UML, que é o formalismo padrão para análise, modelagem e *design* de sistemas com OO.

Percebe-se, destarte, que as abordagens semânticas composicionais procuram compreender o todo, enquanto as abordagens semânticas léxicas focalizam a descoberta do significado de cada unidade lexical. Nessa última abordagem, por conseguinte, é importante o estudo da interrelação entre as palavras, uma vez que as expressões linguísticas não são isoladas. Goddard e Schalley (2010) comentam que as relações semânticas entre elementos léxicos são a base para as *wordnets*. Além disso, as várias aplicações possíveis, dentre elas recuperação e extração de informação, sumarização ou desambiguação,

demonstram a grande demanda que há para resultados na área de relações léxicas e ontologias linguisticamente informadas, mormente quanto à utilização daquelas para construção dessas (HIRST, 2009). A interface entre ontologias e estrutura linguística tem se tornado uma área promissora para representação do conhecimento e PLN, exatamente onde essa pesquisa deseja se enquadrar.

As relações entre elementos léxicos se dividem tradicionalmente em relações horizontais e verticais. Aquelas incluem a sinonímia, palavras diferentes com o mesmo sentido, e a antonímia, palavras diferentes com sentidos precisamente opostos, por exemplo. Jurafsky e Martin (2008) afirmam que aplicações de PLN normalmente estabelecem uma relação mais solta do que a sinonímia, chamada similaridade, e estabelecem métodos estocásticos para cálculo da proximidade. As verticais, por outro lado, são a hiponímia e a meronímia. A hiponímia ocorre quando um elemento léxico chamado hipônimo é mais específico que o hiperônimo. Os exemplos entre ‘automóvel’ e ‘meio de transporte’, ou ‘maçã’ e ‘fruta’ ilustram o conceito. Goddard e Schalley (2010) declaram que, apesar do senso comum, não está correta a presunção de que a hiponímia corresponda à relação taxonômica de tipo. Já a meronímia trata de um elemento, merônimo, que é parte de outro, holônimo. ‘Mão’ e ‘corpo’, ou ‘página’ e ‘livro’ exemplificam a meronímia. A troponímia é um tipo de hiponímia voltada para verbos, no sentido de um verbo especificar a forma como a ação do verbo hiperônimo é realizada. Um exemplo pode ser ‘mover’ e ‘esgueirar’, ou ‘comunicar’ e ‘vociferar’. A implicação, por fim, também é exclusiva para verbos e representa um evento que implica outro, como ‘ressonar’ e ‘dormir’, por modelo.

Cruse (2011) define essas mesmas relações, porém com uma classificação diferente. As relações paradigmáticas de inclusão e identidade são a sinonímia, a hiponímia e a meronímia. Ele ainda acrescenta a taxonímia como um subtipo da hiponímia. Já as relações paradigmáticas de exclusão ou oposição são a incompatibilidade e a antonímia, entre diversas outras. Também discute a diferença das hierarquias lexicais. A hierarquia taxonômica é essencialmente um sistema e classificação, com níveis muito bem definidos, onde é formalizada a experiência no mundo real. Já a hierarquia merônima é uma segmentação de todo-parte.

Ainda na interrelação entre unidades léxicas, um rápido comentário sobre papéis semânticos deve ser feito. Os papéis são assumidos pelas entidades do texto em relações específicas por situação, diferentemente das relações ontológicas, que são mais estáveis. Goddard e Schalley (2010) e Jurafsky e Martin (2008) trazem uma tabela de papéis semânticos e suas respectivas descrições, dentre os quais se citam, para exemplificação, o agente (o que deliberada e objetivamente instiga uma ação ou evento), o executante (aquele que executa a ação), o paciente (a coisa que está em determinado estado ou condição ou passa por uma mudança de estado ou condição), a origem, a localidade, o caminho (rota), o objetivo, além de outros.

A pesquisa de Charton e Torres–Moreno (2011) utiliza o conceito de relações semânticas entre elementos léxicos para extração de significado de textos. Os autores procuram modelar, por meio de um algoritmo simples, os elementos do texto que são ligados por conectores lógicos. O objetivo é decodificar relações de causa, consequência, oposição ou adição, por exemplo, montando um AEF que permita, em última instância, a substituição de um conectivo lógico por outro, ou de unidades lexicais sinônimas, mantendo o sentido original da frase. O tratamento da ambiguidade apresentou grande complexidade uma vez que as palavras utilizadas nos conectores são ambíguas, e normalmente esses conectores são extremamente sensíveis ao contexto de sua utilização. Os resultados experimentais demonstraram-se satisfatórios, atingindo teto de 88% (oitenta e oito por cento) de acerto em operações de substituição, demonstrando a possibilidade de geração de orações completamente novas e diferentes das originais, porém com o mesmo sentido. As aplicações na área de sumarização de textos ou arquiteturas de pergunta e resposta são promissoras.

Já Oliveira, Santos e Gomes (2010) também discutem as relações semânticas entre palavras. O trabalho, em língua portuguesa, apresenta um recurso lexical constituído por afinidades entre termos as quais são extraídas automaticamente de um dicionário. O objetivo é construir uma ontologia lexical assistida por computador a partir do conhecimento esquematizado em um dicionário geral. Os autores definem ontologia lexical como uma estrutura de conhecimento que relaciona unidades lexicais de uma linguagem por meio das relações semânticas entre elas. Além disso, essa ontologia não é uma ontologia de domínio,

a qual se restringe a campos específicos, todavia um esquema de representação que pretende abranger toda a língua. Para atingir seus objetivos, a pesquisa cria gramáticas semânticas para cada tipo de relação que se deseja extrair, e um processo automático extrai, inspeciona e ajusta as ligações detectadas. Esse é um resultado fundamental para esta pesquisa, em particular, uma vez que o esquema de representação do conhecimento construído encaixa-se na arquitetura proposta para análise semântica.

Uma diferença importante deve ser destacada. As relações semânticas entre elementos léxicos são relações entre palavras, as quais se baseiam em relações ontológicas entre os conceitos que constituem os significados dessas palavras. Cimiano, Völker e Buitelaar (2010) detalham essa diferença, especificando que um BD léxico, ou um tesouro, são objetos linguísticos, enquanto uma ontologia é uma teoria lógica. Até porque conceitos e propriedades são definidos em bases lógicas, contrapondo-se a unidades lexicais, as quais se organizam em um léxico. Guarino (1998) define a relação ontológica como uma relação entre as conceitualizações, ou seja, qualquer relação conceitual entre duas entidades.

Nesse aspecto, Nirenburg e Raskin (2004) ensinam que a semântica ontológica é uma teoria de significação em linguagem natural e uma abordagem para PLN que usa uma ontologia como o recurso central para extração, representação de significado e raciocínio sobre conhecimento derivado de textos em linguagem natural. Acrescentam que o objetivo do trabalho é desenvolver uma teoria semântica geral que seja formal e detalhada o suficiente para suportar PLN por um computador. Assim, a representação do significado de um texto é derivada dos seguintes passos, os quais incorporam informação que é usualmente delegada à análise pragmática:

- Estabelecimento do significado de cada unidade lexical do texto;
- Estabelecimento do significado das orações do texto;
- Desambiguação desses significados;
- Combinação desses significados em uma estrutura semântica de dependência que cubra:

- O conteúdo semântico proposicional, incluindo causalidade, temporalidade e outras relações entre declarações;
 - As atitudes do autor ou protagonista em relação ao conteúdo proposicional;
 - Os parâmetros da situação fática do discurso.
- Preenchimento das lacunas da estrutura embasado no conhecimento instanciado da própria estrutura, assim como no conhecimento ontológico.

Note-se que a relevante questão da ambiguidade existe no escopo da linguagem, porém nunca no escopo da ontologia. Cimiano, Völker e Buitelaar (2010) defendem que o conceito nunca é ambíguo, o que pode oferecer mais de um sentido é a unidade léxica selecionada para materializar o conceito na linguagem natural. Por isso a ontologia pode ser o diferencial para análise semântica em PLN, com aplicações que interessam a este estudo, qual seja a IA, ou até no campo de TA, considerando que a ontologia não deve ser específica a uma linguagem, porquanto ela modela a existência. Esse é o marco teórico desta pesquisa.

Essa ontologia, comportando-se como fonte de conhecimento estático para o sistema de PLN, inclui alguns modelos. Nirenburg e Raskin (2004) os listam como um modelo do mundo físico, um modelo do discurso dos participantes, incluindo seus objetivos e atitudes para com os elementos da ontologia, e conhecimento sobre a situação de comunicação que se apresenta. Além disso, um repositório de fatos contendo instâncias lembradas de eventos e objetos é necessário. O significado sentencial, portanto, é definido como uma expressão – representação do entendimento textual – obtida por meio da aplicação do conjunto de regras de análise sintática do texto para ligação das dependências sintáticas às dependências ontológicas, estabelecendo, por extensão, o significado de cada unidade lexical.

A pesquisa de Shinde, Bhojane e Mahajan (2012) é uma aplicação direta desse princípio. Os autores utilizam um motor de PLN para extrair de textos livres em linguagem natural inglesa, contendo especificação de requisitos de *software* de usuário, diagramas de classes da UML. Para isso utilizam a análise morfológica para

lematização e extração de verbos e substantivos. Após esse passo, a árvore de derivação da análise sintática é ligada a uma ontologia, no caso a *wordnet*, para extração de relações semânticas entre as classes e atributos. Uma crítica à pesquisa cabe no momento em que os autores consideram a *wordnet* uma ontologia, o que não é completamente correto considerando-se que aquela é um objeto linguístico, enquanto uma ontologia é uma entidade maior, como já discutido. Por fim, os autores ainda implementam um módulo de codificação automatizado para conversão do diagrama de classes em código fonte de programação.

Joshi e Deshpande (2012) também aplicam diretamente o mesmo princípio. Em uma pesquisa bastante parecida com a de Shinde, Bhojane e Mahajan (2012), tanto nos objetivos quanto na execução, eles propõem um diferencial, qual seja a utilização de uma ontologia de domínio para extração de interpretação semântica. A *wordnet* é utilizada exclusivamente nas análises morfológica e sintática. Esse é um modelo bastante mais robusto que o outro, e é precisamente na semântica ontológica que esta pesquisa, em particular, vai embasar seus pressupostos.

Mudando o foco da discussão, outro método para análise semântica em PLN é a utilização de grafos. Sowa (1979) escreve o artigo seminal onde descreve os grafos conceituais, quais sejam uma linguagem para representação de conhecimento e um padrão para construção de modelos. Um grafo conceitual, portanto, é uma representação de conceitos e relações entre eles, como uma rede semântica, onde axiomas e regras de formação são estruturados.

É necessário, entretanto, definir precisamente o que seja um grafo. Mihalcea e Radev (2011) o definem como uma estrutura de dados que contem um conjunto de vértices conectados por um conjunto de arestas, que podem ser utilizados para modelar relacionamentos entre objetos de uma coleção. Eles normalmente são estudados na teoria dos grafos, área de pesquisa da matemática. O problema clássico que instiga a criação da teoria dos grafos são as sete pontes de Königsberg. A Figura 6 ilustra, à esquerda, o rio que forma ilhas na cidade. Para transitar entre essas regiões, é necessário cruzar as pontes, em vermelho. A questão se apresenta por ser ou não possível atravessar todas as pontes passando

apenas uma única vez por cada uma delas. Euler, em 1741, demonstra que não é possível, e a investigação da teoria dos grafos inicia-se. O desenho à direita apresenta o grafo que modela o problema.

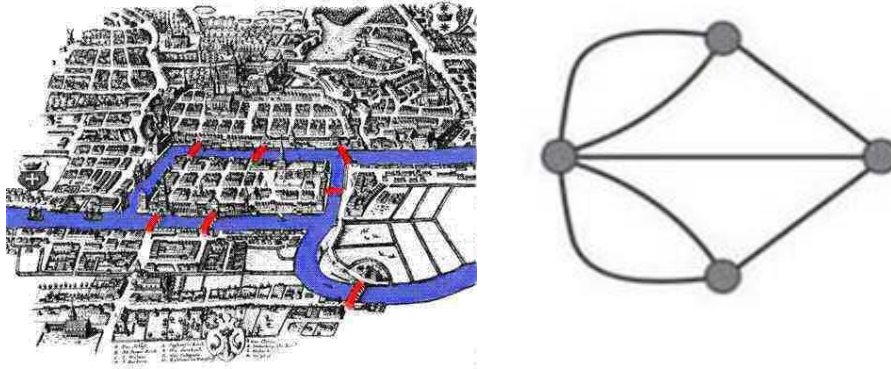


Figura 6 – As pontes de Königsberg (ARAÚJO, 2012).

Os grafos são uma estrutura de dados de tremenda importância para a CC. As utilizações são incontáveis, nos mais diversos campos, tais como algoritmos para roteamento de redes de computadores, modelagens de BD distribuídos, e de especial interesse desta pesquisa, PLN. São classificados como orientados ou não orientados, dependendo do sentido das arestas. Podem ser representados por meio de matrizes, onde os vértices são descritos nas linhas e colunas, enquanto as intersecções abrigam os pesos das arestas, ou então podem ser utilizadas listas encadeadas para representação. Vários algoritmos existem na teoria dos grafos úteis para PLN, tais como navegação em profundidade, conversão em árvore, percorrimento mais curto, corte, verificação de isomorfismo, redução dimensional, e processos estocásticos (MIHALCEA; RADEV, 2011).

Tanto a teoria dos grafos quanto a área de PLN são disciplinas bastante estudadas, porém com algoritmos, aplicações ou potenciais usuários finais diferentes. Diversas pesquisas têm mostrado, no entanto, que as disciplinas são de fato intimamente conectadas, com aplicações de PLN e RI se beneficiando diretamente dos *frameworks* da teoria dos grafos. A pesquisa de Heinecke *et al.* (2008), por exemplo, usa um grafo para modelagem semântica e depois constrói uma ontologia com as unidades lexicais selecionadas. Já Widdows e Dorow (2002) apresentam um método não supervisionado de reconhecimento semântico sobre etiquetagem POS usando grafos. Nela, as relações sintáticas reconhecidas pelo

etiquetador POS são modeladas como arestas de um grafo, com as unidades lexicais e respectivas etiquetas nos vértices. Sinha e Mihalcea (2007), por fim, propõem outro método não supervisionado para desambiguação de sentido usando grafos. O objetivo é utilizar algoritmos de centralização de vértices para contextualizar as palavras. Relembre-se que os métodos supervisionados oferecem resultados melhores, entretanto têm o gargalo de obrigar a geração de *corpus* anotado manualmente para aprendizado do ferramental. Pesquisas como essas, destarte, que introduzem novas abordagens não supervisionadas com desempenho equivalente podem efetivamente ser o futuro de PLN, e a teoria dos grafos tem potencial para suportar isso.

Concluindo, portanto, a análise semântica procura aprofundar a análise que, na grande maioria das pesquisas de PLN, fica no nível sintático. Goddard e Schalley (2010) trazem à discussão um aspecto interessante que diz respeito à questão da padronização. Conquanto não seja possível, nem talvez desejável, a padronização da metodologia de trabalho dos inúmeros grupos de pesquisa e comunidades envolvidas com PLN, mormente em línguas diferentes, ainda assim uma atenção deve ser dada à possibilidade de comparação dos resultados de pesquisa e interoperabilidade dos sistemas. Só assim a área vai crescer e se estabelecer como ciência. A ainda grande utilização do cálculo de predicados da lógica de primeira ordem em PLN, não obstante suas deficiências, é um claro exemplo do valor que tem uma notação padrão largamente compreendida e aceita. Talvez Schalley (2004) esteja certa em usar um padrão como a UML, e nisso haja potencial para tratar essa questão.

2.2.5. ANÁLISE PRAGMÁTICA

A análise pragmática é aquela onde o contexto e o discurso são perscrutados. Cherpas (1992) a define como o trabalho de análise funcional de uma linguagem. Comenta que do ponto de vista da semântica, o objetivo é estabelecer o que determinada sentença declara. Do ponto de vista da pragmática, por outro lado, a sentença nada declara: quem declara são as pessoas, e elas o fazem por alguma razão ou objetivo. Isso deve ser primordialmente descoberto. Um sistema de PLN

nunca vai ser completamente efetivo enquanto considerar o texto como um conjunto de elementos linguísticos independentes entre si e de seu autor. Já Sabah (2011) propõe que a compreensão do significado superficial de uma linguagem natural não é suficiente, no entanto os objetivos, intensões e estratégias dos participantes em um diálogo devem ser entendidos. Alega contundentemente que o entendimento não é somente embasado em critérios lógicos, mas também o resultado emergente de processos não racionais que não podem ser descritos por meio de um algoritmo matemático.

A exploração se dá por meio do reconhecimento de que as sentenças de um texto não são isoladas, mas sim relacionadas e significativas entre si, formando um discurso. Moens, Uyttendaele e Dumortier (1999) alegam que a análise do discurso provê conhecimento valioso a ser incorporado a sistemas de processamento de textos. Ainda acrescentam que, em tarefas que requeiram entendimento textual, o conhecimento de padrões do discurso que sejam comuns a determinados tipos de textos é muito útil. Mitkov (2010) afirma que o que se espera do discurso é coesão e coerência. A coesão se apresenta pela forma como as unidades textuais são ligadas. Ela acontece quando a interpretação de algum elemento do discurso dependa de outro e envolva o uso de alguma alternativa linguística que se refira ou substitua itens mencionados anteriormente, por meio de anáforas, por exemplo. Já a coerência trata das relações de significado entre duas unidades, e como duas ou mais unidades se combinam para produzir o sentido geral da declaração.

Em relação à coesão, o discurso é estruturado, destarte, por meio de uma unidade maior que uma oração, como um parágrafo, episódio ou tópico. Hearst (1994) constrói um algoritmo capaz de particionar textos expositivos em unidades de discurso coerentes, compostas de vários parágrafos, os quais refletem a estrutura de tópicos do texto. Esse é um algoritmo de abordagem estatística que usa a frequência léxica de coocorrência e distribuição de informação para reconhecer as interações entre múltiplos temas simultâneos. Hearst (1997) demonstra como a modelagem é útil para RI e sumarização automática e discute que seus resultados de pesquisa correspondem bem ao julgamento da estruturação textual realizado por pessoas.

Choi, Wiemer–Hastings e Moore (2001), por sua vez, evoluem esse mesmo algoritmo com a aplicação de análise semântica latente (ASL) conseguindo resultados significativamente melhores. A ASL é uma abordagem de classificação para expansão de parâmetros de pesquisa. Ela preconiza que o significado de uma unidade lexical é representado pela sua relação com outras unidades. Uma matriz de similaridade é, então, montada e uma análise vetorial é realizada para distinguir palavras diferentes.

Já em relação à coerência, Hobbs (1979) propõe a criação de relações de coerência, tais como causa, avaliação, paralelismo, elaboração, entre outras. É proposto um mecanismo de inferência para raciocínio sobre essas relações, com uma forma de representação onde se pode aplicar lógica de primeira ordem. É um modelo muito bem utilizado para solução automática de anáforas. Mitkov (2010) também apresenta a teoria de estrutura retórica, muito adotada por pesquisadores de PLN, onde as relações são desenhadas entre unidades de texto denominadas núcleo e satélites. O núcleo é a unidade principal, que pode ser interpretada isoladamente. Os satélites se conectam a ele por meio de relações. Uma delas é a prova, que descreve a evidência que demonstra a alegação nuclear. Outra é a circunstância, que contextualiza o núcleo de forma que o mesmo possa ser interpretado. Há várias outras sendo incorporadas ao modelo ao longo da experiência acumulada das pesquisas na área.

Cruse (2011) classifica os atos do discurso. Ele seleciona verbos que são representantes dos seguintes atos: assertivos, diretivos, comissivos, expressivos e declarativos. Os primeiros comprometem o agente à verdade expressada pela proposição: sugerir, declarar, reportar. Os diretivos intencionam o início de uma ação por parte do paciente: ordenar, comandar, requerer. Os comissivos comprometem o agente a uma ação futura: prometer, ofertar, contratar. Os expressivos retratam a atitude psicológica do agente em relação a um estado atual: agradecer, perdoar, blasfemar. Os últimos, por fim, causam uma mudança na realidade e codificam essa mudança tais quais suas consequências: resignar, divorciar, sentenciar. Apesar de reconhecer que a classificação dos atos do discurso é útil para análise pragmática, Cruse (2011) percebe que não há razões para acreditar que esse seja um conjunto

finito, ou seja, vários outros atos existem e a classificação depende diretamente da aplicação.

A análise pragmática é a mais imatura em PLN. Sua profundidade e complexidade intimidam pesquisas na área. Percebe-se, porém, que aplicações podem bem utilizar seus resultados, sobretudo aquelas que mais se aproximam do usuário final, quais sejam motores de TA ou sistemas de pergunta–resposta. Esta pesquisa, em particular, limita-se à análise semântica, não percebendo vantagem estratégica para a IA com fins de RI em tal nível de profundidade linguística.

Concluindo o referencial teórico de PLN, o mesmo se associa a esta tese para cumprimento de seu primeiro objetivo específico. O levantamento colabora para uma área de pesquisa que ainda demanda estudos aprofundados e resultados abrangentes, qual seja a análise semântica em língua portuguesa, o que é uma das contribuições deste trabalho. Além disso, o desenvolvimento e integração de ferramental para sistematização do PLN é um processo possível a partir dos pressupostos observados e dos resultados de investigações prévias elencados.

2.3. RECUPERAÇÃO DE INFORMAÇÃO

A área de RI pode ser vista, sob alguns aspectos, como uma aplicação de sucesso de PLN. O crescimento rápido, desordenado e acachapante da Internet só foi possível por causa de motores de busca livres, disponíveis e efetivos, a maioria desenvolvida com técnicas de PLN. Savoy e Gaussier (2010) estimam que 85% (oitenta e cinco por cento) dos usuários de Internet usam essas ferramentas quando pesquisam por alguma informação específica. Baeza–Yates e Ribeiro–Neto (2011) definem RI como a disciplina que se ocupa com a representação, armazenamento, organização e acesso a itens de informação.

Jansen e Rieh (2010) discutem a diferença entre as áreas de pesquisa de informação e RI. Embora ambas tratem da interação entre pessoas e conteúdo digital em SI, a primeira focaliza especificamente esse intercâmbio entre o usuário e o sistema de recuperação. O amplo campo de atuação vai desde o estudo do

comportamento das pessoas quanto à localização da informação, à adoção de estratégias de pesquisa, até ao julgamento da relevância da informação recuperada. Já a RI, em contraste, se preocupa com o encontro de material de natureza não estruturada, armazenado em grandes coleções digitais de quaisquer formatos, que satisfaça uma necessidade de informação. A RI é a extração de informação de uma coleção de conteúdo. Por isso sua proximidade com PLN. Os autores ainda atestam que, tradicionalmente, a pesquisa de informação é uma área de cientistas da informação e bibliotecários, enquanto a RI é uma área de cientistas da computação. Argumentam, contudo, que atualmente as áreas têm se correlacionado, inclusive com resultados de uma sendo reaproveitados na outra, e, de principal importância, com pesquisadores em influência mútua.

Já Baeza–Yates e Ribeiro–Neto (2011) diferenciam a RI da recuperação de dados (RD). Enquanto a primeira se preocupa com extração de informação a partir de conteúdo não estruturado, a última focaliza especificamente a determinação de quais documentos de uma base contém as palavras chaves de uma pesquisa. Isso, entretanto, frequentemente não atende à necessidade de informação do usuário. Na RI, o usuário espera, ou pelo menos acha desejável, que documentos que contenham sinônimos dos parâmetros de pesquisa sejam retornados, enquanto na RD isso não ocorre. Percebe-se que, na RI, erros poluem o resultado de pesquisa, porém, se em pequena escala, podem passar despercebidos. Na RD, por outro lado, um único objeto recuperado erroneamente dentro de milhões de registros representa total fracasso da aplicação. Os BD relacionais bem exemplificam a questão.

Um dos maiores problemas da RI é o fato de que os sistemas de recuperação têm de lidar com descrições imprecisas e incompletas tanto do lado de quem pesquisa, quanto do lado do repositório de informações (SAVOY; GAUSSIER, 2010). Os usuários usam linguagem não padronizada e imprecisa para submeter consultas aos sistemas, e esses possuem informação parametrizada incompleta dos documentos em suas bases. Isso contrasta com a situação da área de BD. Nela as informações encontram-se armazenadas de forma controlada e normalizada. Além disso, a linguagem utilizada para submissão de pesquisas é precisa e não ambígua: a linguagem estruturada de consulta (SQL). Já em uma pesquisa a um motor de

busca da Internet, o usuário usa descrições ambíguas e curtas para o que efetivamente deseja, e o sistema utiliza uma abordagem de solução de problema via tentativa e erro para encontrar a informação, ao invés da abordagem ideal que seria o paradigma de pergunta e resposta. Por último, uma pesquisa SQL sempre devolve um resultado determinístico, enquanto uma pesquisa a sistema de RI retorna o conjunto de melhores respostas possíveis ordenadas por probabilidade de relevância à consulta submetida.

Além desse problema, Savoy e Gaussier (2010) acrescentam ainda as idiossincrasias da linguagem natural, estudadas na Seção 2.2.4, tais como a homonímia, polissemia ou sinonímia, por exemplo. A estratégia básica para RI, qual seja a extração de palavras dos documentos e comparação dessas com as oferecidas nas consultas, não vai gerar resultados satisfatórios se aqueles fatores forem ignorados. Por isso o PLN é intrinsecamente ligado à RI, e esta pesquisa se insere nessa discussão.

Há vários modelos para RI. Baeza–Yates e Ribeiro–Neto (2011) argumentam que a modelagem de sistemas de RI é um processo complexo que objetiva primordialmente a produção de uma função de ranqueamento, ou seja, uma função que associe uma pontuação a cada documento de uma base de dados em relação à determinada pesquisa. Assim, uma nota de corte pode ser estabelecida para definir quais documentos serão recuperados, e a ordem dos mesmos será instituída pela pontuação decrescentemente. O processo constitui-se de duas tarefas principais. A primeira é a concepção de um *framework* lógico para representação dos documentos e dos parâmetros de pesquisa. O segundo é a definição daquela função de ranqueamento. O *framework* lógico é normalmente embasado em lógica booleana, vetores, ou distribuição probabilística, os quais serão discutidos mais detalhadamente a seguir. Em resumo, os autores formalizam a definição de um modelo de RI por meio da tupla abaixo:

$$M = [D, Q, F, R(q_i, d_j)]$$

Onde:

M : modelo de RI.

D : conjunto de representações dos documentos de uma coleção.

Q : conjunto de representações das necessidades de informação do usuário. Usualmente as necessidades de informação são transcritas, ou representadas, por parâmetros de pesquisa.

F : *framework* lógico para representação dos documentos e dos parâmetros de pesquisas.

$R(q_i, d_j)$: função de ranqueamento que recebe como parâmetros uma representação de parâmetro de pesquisa q_i (membro de Q) e uma representação de documento d_j (membro de D), produzindo um número que será utilizado para pontuação do documento naquela pesquisa.

Baeza–Yates e Ribeiro–Neto (2011) ainda exemplificam que uma representação de documento pode ser um subconjunto de todas as palavras do documento, utilizando-se PLN para remover artigos ou preposições. A representação de necessidade de informação pode ser um exercício de expansão dos parâmetros originais de uma pesquisa com seus respectivos sinônimos. O *framework* que vai consolidar as representações deve prover uma intuição para construção da função de ranqueamento.

O primeiro modelo que se deseja descrever, mais clássico, é o modelo booleano. Esse modelo usa os pressupostos da teoria de conjuntos e da álgebra booleana. Nele os documentos são representados por um conjunto de palavras chaves de indexação, extraídas automaticamente dos próprios documentos ou providas por indexadores, verificadas em um vocabulário controlado (ANDERSON; PÉREZ–CARBALLO, 2001). As consultas são submetidas por meio do uso dos termos de indexação associados via operadores lógicos (e, ou, não). Esse é um modelo muito utilizado que tem uma longa tradição na biblioteconomia e é fácil de ser eficientemente implementado. Possui várias deficiências, contudo, tais como a impossibilidade de se ranquear os documentos, as limitações da lógica binária, a não solução da sinonímia, e a grande dependência da escolha dos descritores para qualidade do processo (SAVOY; GAUSSIER, 2010). Esses problemas podem, evidentemente, ser tratados adicionando-se algumas propriedades. Os modelos mais recentes, todavia, apresentam desempenho geral consistentemente melhor.

A indexação no modelo vetor–espaço já é mais elaborada. Savoy e Gaussier (2010) colocam que os descritores selecionados têm pesos, de forma que os usuários não precisam obrigatoriamente utilizar conectivos lógicos para realizarem suas pesquisas, sendo expressões em linguagem natural suficientes. No modelo, as consultas e os documentos são representados por vetores num espaço multidimensional. Cada dimensão corresponde a um descritor e os vetores carregam a importância daquele descritor no documento ou na consulta. O grau de similaridade entre um documento e uma consulta, então, é calculado por meio de fórmulas de peso. Várias fórmulas há na literatura sem se chegar ainda a um consenso de qual delas é a mais efetiva. Uma abordagem interessante é a utilização de ASL (FOX, 2010) no espaço vetorial, a qual permite a derivação automática de informação semântica de uma coleção de documentos a partir da análise de coocorrência. A sinonímia e a polissemia são tratadas, nesse caso. A complexidade computacional, no entanto, é bastante alta.

Os modelos probabilísticos, por outro lado, analisam a recuperação como um processo de classificação. Para cada consulta, o sistema monta duas classes, relevante e não relevante. Assim, a recuperação é calculada pela probabilidade bayesiana de um documento pertencer ou não à classe, com sua respectiva pontuação para ordenação. Savoy e Gaussier (2010) ainda apõem que podem ser utilizadas novas variáveis para fortalecer o cálculo probabilístico, tais como frequência de termos, frequência de documentos e tamanho do documento. A divergência de aleatoriedade usa essas variáveis em uma interessante formulação que analisa a distribuição de frequência de um determinado termo em um documento e o risco do mesmo ser utilizado como descritor considerando sua ocorrência em todos os outros documentos.

Savoy e Gaussier (2010) percebem que, de maneira geral, todos os modelos de RI têm suas vantagens e desvantagens. Cada um deles tem resultado ótimo, ou seja, aquele em que toda a potencialidade do modelo é explorada, em um conjunto definido de aplicação. Todos, entretanto, apresentam desempenho similar considerando o caso médio. Baeza–Yates e Ribeiro–Neto (2011) defendem que o modelo booleano é considerado o mais frágil deles, enquanto há certa controvérsia

na literatura entre qual dos modelos vetoriais ou probabilísticos tem melhor desempenho.

Uma tentativa de melhorar os resultados é a produção de modelos híbridos, denominados fusão de dados, os quais procuram aproveitar os pontos fortes de diferentes abordagens. A grande desvantagem é a enorme complexidade computacional, o que obriga o aumento de espaço de armazenamento e o tempo de processamento. Uma estratégia com potencial é a utilização do conceito de acoplamento bibliográfico para páginas da Internet. O acoplamento bibliográfico procura medir similaridade de assunto por meio de referências a documentos. Se o mesmo conceito for aplicado a *hyperlinks* de páginas da Internet, podem-se construir agrupamentos que interessam a sistemas de RI.

A pesquisa de Yuan e Belkin (2010) é um exemplo de abordagem híbrida. A proposta é combinar diferentes modelos de RI, mormente quanto a suas técnicas de suporte, de várias formas para produzir quatro arranjos diferentes. Os autores citam algumas técnicas de suporte, tais como melhor correspondência ou correspondência exata, para comparação. Acrescentam indexação, automática ou manual, ou classificação, para representação. Para apresentação, sugerem agrupamento ou listagem. Cada uma dessas combinações é avaliada quanto aos resultados de RI para estratégias de procura de informação específicas, comparando-se a sistemas estáticos. Percebe-se grande vantagem, promovendo a proposta de um *framework* de técnicas de RI que possam ser combinadas em tempo real para realização de pesquisas em ambientes heterogêneos.

Para concluir a discussão dos modelos de RI, menciona-se brevemente uma estratégia para melhorar a combinação entre a consulta e os documentos retornados. Ela se baseia na expansão dos parâmetros de pesquisa. O princípio geral é aumentar os parâmetros usando palavras ou expressões similares ou semanticamente relacionadas àquelas selecionadas pelo usuário. Para isso pode-se utilizar um tesouro, um esquema de representação do conhecimento mais robusto ou derivar a informação de uma coleção. Savoy e Gaussier (2010) mencionam que esse processo, que pode ser totalmente automático, pode também ser iterativo, com o sistema sugerindo acréscimos às consultas feitas pelo usuário para que ele

mesmo selecione quais sugestões deseja acatar antes de submeter os parâmetros à pesquisa.

Uma aplicação inspirada neste modelo é realizada por Wu, Zhang e Liu (2012). Em sua pesquisa, eles propõem um mecanismo automático com treinamento não supervisionado para segmentação de parâmetros de pesquisa. O objetivo é agregar a sequência de unidades lexicais em compostos complexos, com semântica mais significativa, para melhorar os resultados dos sistemas de RI. Para isso, é proposto um novo algoritmo embasado nos pressupostos teóricos dos sistemas de TA com abordagem estatística. Além de apresentar resultado experimental satisfatório, o grande diferencial desse trabalho é a proposta de um método independente da linguagem, com treinamento não supervisionado, o que permite sua migração para qualquer domínio com muita facilidade.

Já a pesquisa de Blanco e Lioma (2012) é uma interessante aplicação de PLN a RI. Eles argumentam que a abordagem padrão para RI é considerar o texto como um conjunto de palavras, e realizar tratamentos para identificação e mapeamento desses termos aos parâmetros da pesquisa. Uma alternativa, entretanto, é considerar o texto como um grafo. Essa proposta pode ser materializada modelando-se as unidades lexicais como vértices do grafo, cujas arestas sejam representações de coocorrência ou relações gramaticais entre as palavras. Assim, os algoritmos matemáticos de navegação em grafos podem ser aplicados a esse modelo para medição de várias propriedades importantes do grafo, e conseqüentemente do texto original. Com essa estratégia, foi possível integrar elementos do discurso na análise textual, tais como coerência, fluxo e densidade, durante a recuperação.

Uma questão preponderante em RI, noutra prisma, são as medidas de avaliação. Baeza–Yates e Ribeiro–Neto (2011) explicam que avaliar um sistema de RI é medir quão bem o sistema atende às necessidades de informação de seus usuários. Isso é naturalmente problemático, todavia, considerando-se que um mesmo resultado de pesquisa pode ser interpretado diferentemente por usuários distintos. Ainda assim é possível definir algumas métricas aproximadas que, de maneira geral, têm alguma relação com as preferências de uma população de

usuários. A avaliação de sistemas de RI, além de ser fundamental para medida da qualidade do sistema, é também importante para comparação objetiva de resultados de sistemas diferentes, permitindo evolução da pesquisa na área. Concluem cunhando a definição de que avaliação de RI é o processo de associação sistemática de uma métrica quantitativa aos resultados produzidos por um sistema de RI em resposta a um conjunto de pesquisas de usuários. Essa métrica, evidentemente, deve estar diretamente associada à relevância do resultado da pesquisa para o usuário.

Bird, Klein e Loper (2009) detalham os conceitos de precisão e revocação. Para eles, precisão é o indicador de quantos dos itens identificados são relevantes. Já revocação é o apontador de quantos dos itens relevantes foram identificados. Os autores ainda mencionam que os documentos relevantes recuperados representam o verdadeiro-positivo da operação de busca. Isso significa que documentos de interesse foram efetivamente retornados. Já os documentos irrelevantes que foram erroneamente recuperados representam o falso-positivo da operação, ou seja, documentos que não deveriam ser devolvidos, mas o foram. Já os documentos que simbolizam o falso-negativo são aqueles que são relevantes, no entanto não foram recuperados pela pesquisa por alguma espécie de erro. Os documentos irrelevantes para a busca, que não foram retornados pelo motor, por fim, indicam o verdadeiro-negativo. Percebe-se que, em uma abordagem lógica, melhor é o motor de busca quando consegue maximizar as proposições verdadeiras, o verdadeiro-positivo e o verdadeiro-negativo, enquanto minimiza as proposições falsas, o falso-positivo e o falso-negativo. Camara Junior (2007) mostra a Figura 7 para ilustração das formulações de precisão e revocação.

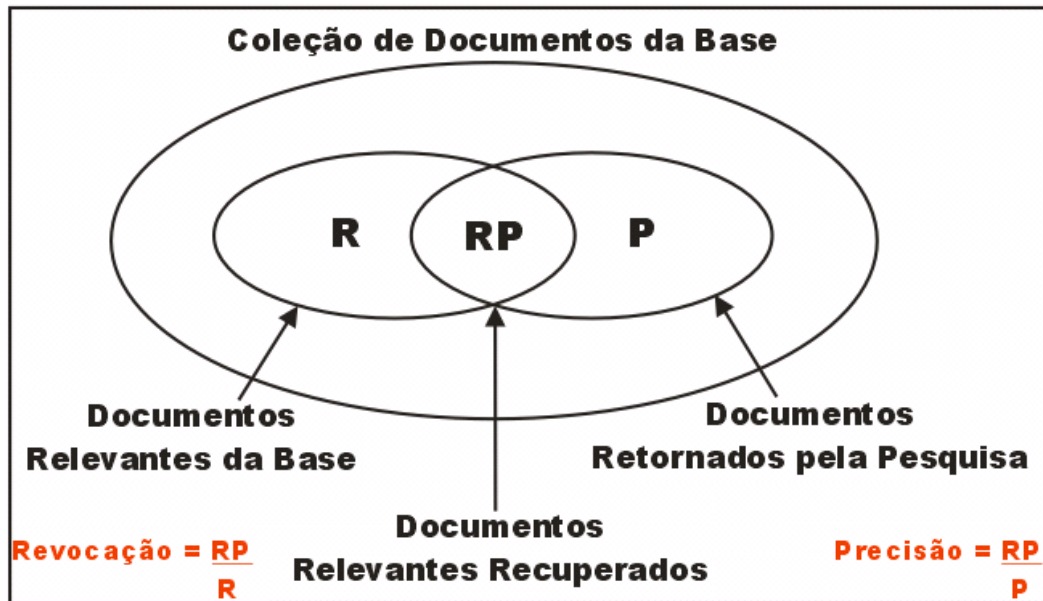


Figura 7 – Definições de revocação e precisão (CAMARA JUNIOR, 2007).

Associando as definições, nota-se que o conjunto RP é o conjunto de verdadeiro-positivo. Esse é o conjunto que se deseja maximizar. O conjunto R – RP, ou seja, os documentos relevantes que não foram recuperados são falso-negativo. O conjunto P – RP, os documentos irrelevantes retornados pela pesquisa, são o falso – positivo. O conjunto de documentos não relevantes e não recuperados aponta o verdadeiro-negativo, saldo que também se deseja elevar ao máximo. O resultado de pesquisa ideal, conclui-se, é aquele onde o conjunto P se sobrepõe exatamente sobre o conjunto R.

Savoy e Gaussier (2010) acrescentam ilustrativamente que o clássico juramento de testemunhas em juris populares é útil para explicar os conceitos: ‘a verdade, toda a verdade (revocação), e nada mais do que a verdade (precisão)’. Ainda definem uma média harmônica entre as duas grandezas, muito utilizada em aplicações de PLN, para gerar um índice de medição concentrado, denominado índice F ou pontuação F, qual seja:

$$F = \frac{2 * IP * IR}{IP + IR}$$

Onde:

F : pontuação F, média harmônica dos índices de revocação e precisão

IP : índice de precisão

IR : índice de revocação

Com a substituição dos parâmetros das definições dos índices de revocação e precisão, ilustrados na Figura 7, além da simplificação matemática dos termos, chega-se à seguinte fórmula, que é aplicada nesta pesquisa:

$$F = \frac{2 * RP}{P + R}$$

Onde:

F : índice F

RP : quantidade de documentos relevantes recuperados pela pesquisa

P : quantidade total de documentos recuperados pela pesquisa

R : quantidade total de documentos relevantes na base de dados

Como uma alternativa, contudo, a avaliação também deveria considerar a ordem que os documentos são retornados na consulta, procurando analisar se os mais relevantes estão efetivamente nas posições topo. Järvelin e Kekäläinen (2002) propõem uma nova medida chamada ganho cumulativo descontado normalizado que é muito bem adaptada para essa necessidade. Com um somatório logarítmico e uma grandeza de normalização, os resultados experimentais sobre uma base de dados da sétima conferência de recuperação de textos (TREC-7) são bastante satisfatórios.

Baeza-Yates e Ribeiro-Neto (2011), embora reconheçam que as medidas de precisão e revocação têm sido extensivamente utilizadas para julgamento da qualidade de algoritmos de RI, fazem algumas críticas aos modelos de avaliação que as utilizam. Primeiramente, comentam que o estabelecimento da revocação é muito caro, sobretudo para bases de dados grandes. Isso se dá porque para calcular a revocação é necessário conhecer todos os documentos relevantes de uma base de dados para uma determinada pesquisa, o que demanda, na verdade, conhecer em considerável nível de detalhe todos os documentos da base. Para coleções numerosas de documentos, esse conhecimento não é possível, ou é

extremamente caro, o que torna o cálculo da revocação impreciso. Nesta pesquisa, em particular, a base de dados utilizada para aplicação dos pressupostos é bastante reduzida, considerando-se que a análise se dá em nível semântico, o que impede a utilização de grandes coletâneas. Em segundo lugar, os autores reconhecem que as medidas de precisão e revocação capturam aspectos muito diferentes do conjunto de documentos recuperados, por vezes contraditórios. Enquanto aquela avalia a pureza do resultado, essa mede sua cobertura. Usar as duas medidas separadamente, portanto, pode não representar adequadamente a qualidade do sistema. Uma combinação dessas medidas, em muitas situações, se apresenta mais apropriada. Nesta pesquisa, opta-se pelo índice F.

Concluindo a discussão da avaliação, Savoy e Gaussier (2010) listam seis categorias de problemas que ocorrem em RI gerando valor 0 (zero) para precisão, os quais precisam ser tratados. A primeira é a lista de palavras ignoradas para indexação, *stopwords*, que deve ser controlada para não excluir unidades lexicais importantes para RI. A segunda trata da lematização, onde um sistema mal calibrado pode agrupar conceitos diferentes num mesmo lema dificultando a recuperação. Já a terceira muda o foco para o lado do demandante, procurando reconhecer os erros de ortografia produzidos pelos usuários na submissão das consultas.

A quarta categoria trata especificamente das idiossincrasias da linguagem, como a sinonímia, por exemplo, e discute que o sistema de RI deve usar técnicas de PLN semântico para correção. A quinta, também do lado do usuário, reconhece a falta de especificidade na produção dos parâmetros de pesquisa, o que de fato é um grande calcanhar de Aquiles para sistemas de RI. A última categoria, por fim, é a habilidade de discriminação, qual seja a capacidade de reconhecer que, embora as palavras selecionadas como parâmetros de pesquisa existam em determinado documento, o mesmo pode não ser relevante para uma consulta. O pior caso é quando um documento desse tipo ainda aparece nas primeiras posições do resultado.

Concluindo o referencial teórico de RI, o mesmo se associa a esta pesquisa no provimento do arcabouço teórico para IA e do mecanismo para

avaliação da arquitetura proposta. O índice F é utilizado para validação da arquitetura com suas vantagens subjacentes de agregar as definições de precisão e revocação em uma única unidade de medida.

2.4. INDEXAÇÃO AUTOMÁTICA

Borko (1977) publica um artigo seminal onde afirma que o papel da CI é explicar, controlar e prever o comportamento da informação. Nesse sentido, o cientista da informação investiga as propriedades da informação, as forças que governam seu fluxo, e as técnicas de processamento para otimização do armazenamento, recuperação e disseminação. Nesse contexto, ressalta que a indexação é uma parte importante dos processos de armazenamento e RI.

Assim, Borko (1977) propõe uma teoria de indexação, a qual explica a natureza da indexação, a estrutura do vocabulário, e a qualidade do índice. Unindo estudos anteriores chega à definição de que um índice é um bem ordenado conjunto de elementos de dados. Um elemento de dado é uma palavra ou termo que possui três propriedades. A primeira é que o elemento de dado tem um significado bem definido. A segunda preconiza que não pode haver decomposição em duas ou mais unidades de significação. A última dita a capacidade do elemento de dado de ser manipulado independentemente de quaisquer outros elementos. Quanto à estrutura do vocabulário, discute o relacionamento entre os termos selecionados para o índice e analisa sua posição terminológica quanto à generalidade/exaustão ou especificidade. Em relação à qualidade, por fim, aposta os conceitos de revocação e precisão para medição da efetividade do índice. Além disso, propõe que o bom índice deve ser capaz de agrupar documentos semelhantes ou relacionados. Esse é um artigo embrionário de tremenda importância para a área de indexação, que conclui afirmando que os pressupostos só podem ser validados por meio de dados empíricos advindos de pesquisa experimental. A comunidade científica pesquisa, até os dias de hoje, a partir dessas hipóteses, e esta pesquisa insere-se neste contexto.

Já Baeza–Yates e Ribeiro–Neto (2011) explicam que o índice é uma estrutura de dados construída em um texto para acelerar pesquisas. Manter e

processar um índice tem complexidade computacional alta, ou seja, é consideravelmente mais difícil do que uma pesquisa sequencial direta, por exemplo. Ainda assim, na maioria dos casos, principalmente quando o tamanho das bases de dados aumenta, é obrigatória sua utilização para obtenção de tempos de resposta aceitáveis. Por isso a indexação é uma área de pesquisa de grande importância para a CI e para a CC.

A indexação, portanto, é o processo por meio do qual se definem os termos que serão utilizados para armazenamento nos índices dos documentos (ROBREDO, 2005). Para Lancaster (2004), o objetivo central da indexação é a representação dos documentos existentes na base para fins de recuperação. Dois tipos de indexação concorrem para essa finalidade. O primeiro é a indexação por extração, qual seja aquela onde os termos utilizados para construção do índice são selecionados dentro do próprio texto. Já o segundo, indexação por atribuição, exaustivamente mais complexo de ser realizado por meios automáticos, utiliza termos que não necessariamente encontram-se no texto para descrevê-lo e indexá-lo.

A finalidade última da indexação é precisamente a RI que satisfaça as necessidades de potenciais usuários. Os usuários mais beneficiados pela indexação são aqueles que almejam dados sobre determinado assunto, mas não conhecem quais documentos da base de dados versam sobre ele. Os usuários que apenas desejam reconhecer documentos por meio de suas características próprias, tais como autor, título, data da publicação, edição, etc., não veem na indexação ferramenta que traga grande diferencial (ROBREDO, 2005).

A Figura 8 apresenta uma visão macro do processo de RI, destacando a importância da indexação no procedimento.

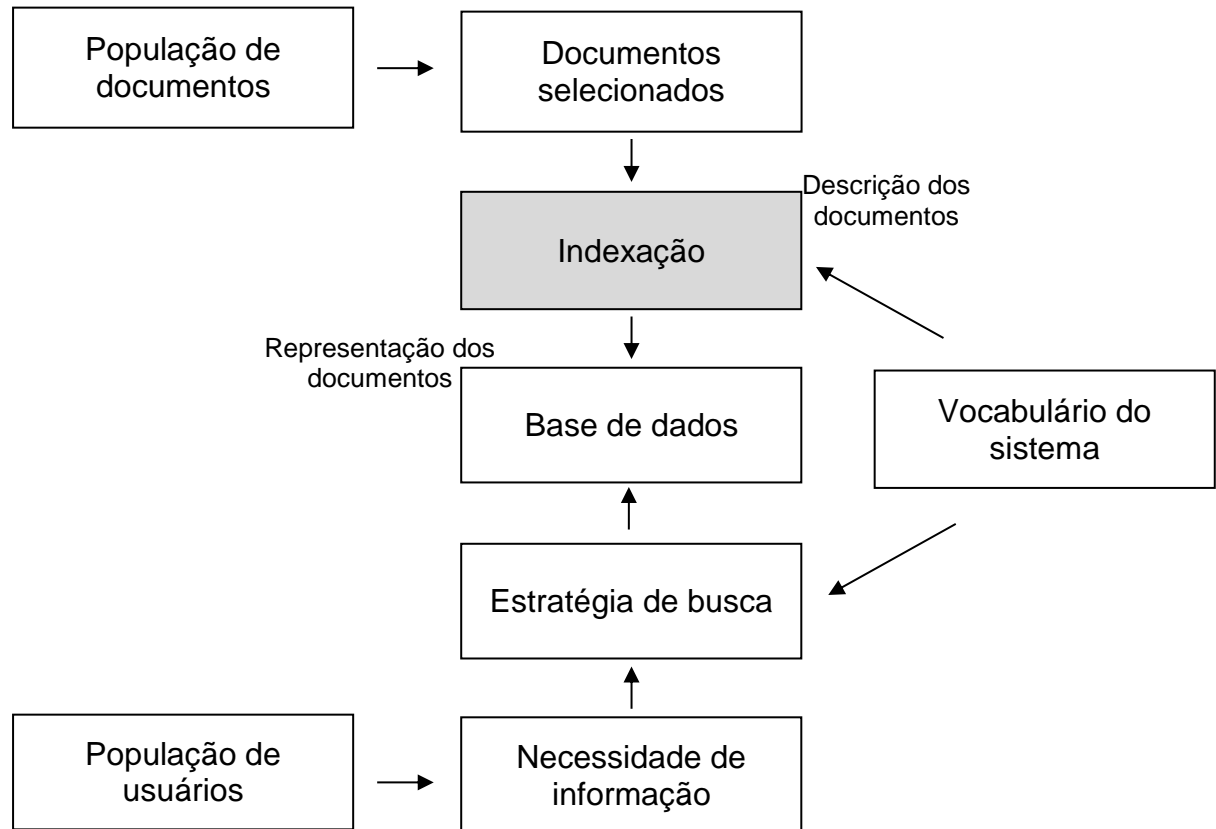


Figura 8 – Quadro amplo da RI (LANCASTER, 2004).

Percebe-se, destarte, que a indexação possibilita a criação de uma interface entre a população de usuários, com suas respectivas necessidades de informação, utilizando uma determinada estratégia de busca, e a base de representação dos documentos. O vocabulário próprio do sistema é o elo que une os sistemas de RI e os BD de indexação. Esse vocabulário é construído por meio de um esquema de representação do conhecimento, utilizado para construção e manutenção do índice. No contexto desta pesquisa é desenvolvida uma ontologia para o módulo semântico de PLN a qual também se aplica para o índice. Expandindo a caixa que representa o processo de indexação propriamente dito, a Figura 9 apresenta um fluxograma que detalha a utilização de um tesauro para validação dos descritores selecionados.

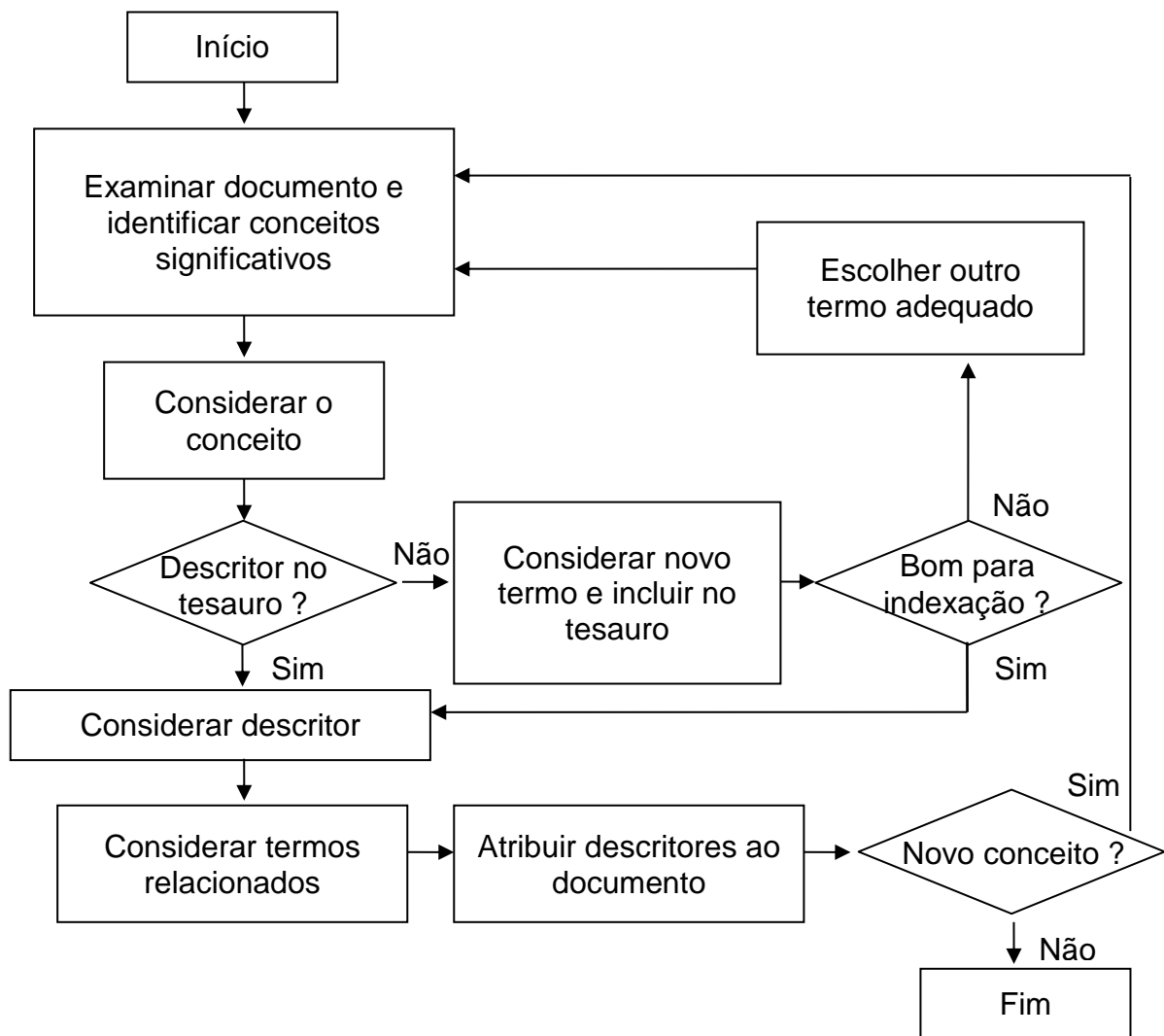


Figura 9 – Fluxograma simplificado do processo de indexação utilizando um tesauro (ROBREDO, 2005).

Lee e Schleyer (2012) demonstram em sua pesquisa a importância da indexação. Considerando as definições já apostas, as quais legam à indexação a função de facilitar o acesso a recursos de informação, os autores perguntam se é possível que algum outro mecanismo facilitador possa apresentar melhores resultados, ou pelo menos resultados equivalentes. Nesse caso, são testadas as etiquetas sociais, ou *folksonomies*, dada sua crescente popularização. Por etiquetagem social os autores entendem o processo dos próprios usuários classificarem e selecionarem descritores para os documentos da base de dados. Ponderando-se o fato de que a indexação é um processo pouco escalável de alto custo, torna-se razoável investigar em qual grau a etiquetagem social poderia

substituir a indexação controlada. Utiliza-se uma base de dados da área de saúde de tamanho expressivo, 231.388 (duzentos e trinta e um mil, trezentos e oitenta e oito) documentos, por sobre a qual há anotações sociais e indexação controlada. A conclusão é que os mecanismos chegam a resultados muito diferentes, tanto no aspecto léxico quanto no semântico, com menos de 1 (um) descritor similar por documento. Isso é usado, na opinião dos autores, para corroborar a tese de que a indexação não pode ser substituída pelas *folksonomies*.

O mecanismo mais básico para indexação de um texto é o índice invertido. Esse é um dispositivo orientado a palavras para indexação de uma coleção de documentos com o objetivo de acelerar as pesquisas. A estrutura é composta de dois elementos, quais sejam o vocabulário e a contagem de ocorrências. Aquele é o conjunto de todas as palavras diferentes dos textos. Para cada uma delas, o índice indica quais documentos as possuem. Essa é a contabilização de quantos documentos possuem cada palavra, e, em cada documento, quantas vezes a palavra ocorre. O índice é chamado invertido porque por meio dele é possível reconstruir os textos originais. Esse é um mecanismo clássico ainda muito utilizado. A implementação computacional é muito eficiente, porque basta uma matriz esparsa para modelar a estrutura de dados, e a navegação nesse tipo de composição é rápida. O revés encontra-se na complexidade espacial, porquanto é preciso um grande espaço de armazenamento para conter esse tipo de índice (BAEZA–YATES; RIBEIRO–NETO, 2011).

Esse mecanismo, entretanto, não é adequado para resposta a pesquisas por frases ou por proximidade, porque ele não tem a informação de onde cada palavra aparece no texto exatamente, como comentam Baeza–Yates e Ribeiro–Neto (2011). Assim, é preciso inserir a posição de cada palavra ou, por extensão, de cada caractere no texto, produzindo um índice invertido completo. Nesse caso, o problema de complexidade espacial aumenta exponencialmente, o que determina a utilização de técnicas de PLN para melhor seleção dos índices, tais como eliminação de preposições e artigos, por exemplo.

Uma vez bem definido o que é indexação, discutidas sua finalidade e importância, e iniciados os mecanismos clássicos de execução, parte-se agora para

a especificação da IA propriamente dita, foco de aplicação desta pesquisa. Williams (2010) faz uma análise histórica discutindo que a invenção da IA usando a abordagem de palavras chaves dentro do contexto deve ser atribuída a Hans Peter Luhn e Herbert Ohlman. Nessa abordagem, palavras chaves do próprio documento são selecionadas para serem utilizadas como índice, ou seja, um mecanismo de indexação por extração. Por contraposição, no enfoque de palavras chaves fora do contexto, a indexação é realizada com termos que não se encontram no documento, uma indexação por atribuição.

Moens (2000) delinea que o processo natural de seleção automática de descritores a partir de textos em linguagem natural é composto de passos, alguns deles fortemente atrelados a resultados de PLN. O primeiro é a identificação das unidades lexicais do texto. O segundo, a remoção de palavras funcionais, artigos e preposições, por exemplo, além das unidades lexicais que possuam grande frequência no domínio dos documentos, as quais não são específicas o suficiente para representar o conteúdo do documento para recuperação. Já o terceiro passo trata da lematização dos descritores selecionados, para otimização do armazenamento computacional. O quarto passo trata da formação de sintagmas, expressões ou composições para serem utilizados como descritores. Enquanto os primeiros passos ficavam no módulo morfológico de PLN, esse passo já exige o módulo sintático, como detalham Constant, Sigogne e Watrin (2012). O quinto focaliza a substituição dos descritores selecionados por suas respectivas classes em um esquema de representação do conhecimento. A autora menciona um tesouro. Esta pesquisa utiliza uma ontologia e o módulo semântico de PLN. O sexto passo, por fim, estabelece os pesos de cada descritor para classificação e definição de sua efetiva utilização no índice ou descarte. Moens (2000) conclui informando que, conquanto esse processo seja padrão e utilizado por muitos sistemas, a ordem dos passos pode ser trocada. Um exemplo é a inversão do segundo passo com o quarto, ou seja, deixar para excluir as palavras não significativas apenas após a formação das expressões.

Baeza–Yates e Ribeiro–Neto (2011) explicam detalhadamente a forma mais tradicional de classificar automaticamente a importância de um termo para ser utilizado como índice de um documento. Embora extremamente simples, ainda

assim esse é o mecanismo fundamental para os esquemas modernos de cálculo de pesos de termos e é utilizado por praticamente todos os sistemas atuais de RI. Nesse caso, utilizam-se duas grandezas quantitativas. A primeira é a frequência do termo no documento. Quanto maior a quantidade de vezes que um termo aparece no texto, mais importante esse termo é para descrever, ou indexar, o documento. Já a segunda é frequência inversa nos documentos. Quanto mais documentos contiverem determinado termo, menos importante esse termo é para descrever qualquer desses textos. Unindo-se os dois conceitos em uma formulação logarítmica denominada frequência do termo–frequência inversa nos documentos (TF–IDF), é possível produzir uma função para cálculo de pesos de termos de uma coleção de escritos e selecionar automaticamente índices adequados para eles.

A indexação semântica latente (ISL) é criada por Furnas *et al.* (1988) como um novo método para IA. Nela, estruturas implícitas de mais alta ordem, ou seja, conceituais, na associação de termos a textos são modeladas para melhorar a detecção de documentos relevantes em pesquisas. É descrita por Baeza–Yates e Ribeiro–Neto (2011) como uma proposta para a recuperação de documentos embasada na indexação de conceitos, e não de palavras. A indexação tradicional, por meio de termos, prejudica a qualidade da RI uma vez que tanto a precisão quanto a revocação são mal influenciadas. Caso um texto relevante não possua determinado termo que foi utilizado na pesquisa, ele não será recuperado, atrapalhando a revocação. Por outro lado, documentos irrelevantes podem fazer parte do conjunto de resultado de pesquisa caso o termo utilizado para procura esteja indexado no texto. Isso é decorrente da relação vaga que há em um processo de recuperação embasado em conjuntos de palavras chaves.

Já a indexação de conceitos considera que o texto em linguagem natural inclui referências a conceitos e relações entre eles. Assim, uma vez extraídos tais conceitos e realizada a indexação por meio deles, um documento pode ser recuperado em uma pesquisa independente de o parâmetro utilizado estar contido nele. Isso implica que um texto pode ser recuperado para determinada pesquisa se compartilhar conceitos com algum outro documento que seja considerado relevante. Para implementar esses pressupostos, a ISL mapeia os vetores de cada texto da coleção e cada parâmetro de pesquisa em um espaço dimensional composto por

conceitos. Percebe-se claramente, portanto, um modelo de RI vetorial. Embora esse *framework* teórico apresente-se promissor, ainda não se concluíram resultados de pesquisas encorajadores utilizando-se ISL (BAEZA–YATES; RIBEIRO–NETO, 2011).

Li e Kwong (2010) concordam atestando que a ISL é bem conhecida para tratamento dos problemas de sinonímia e polissemia na RI. Reconhecem, no entanto, que o desempenho da estratégia se apresenta muito diferente em conjuntos de dados diversos, alguns proporcionando resultados muito bons, e outros revelando resultados decepcionantes. Ainda não foram completamente entendidas quais são as características dos conjuntos de dados que contribuem para essa diferença, e porque elas o fazem.

A pesquisa de Chung, Miksa e Hastings (2010) procura, todavia, alcançar resultados a partir desses pressupostos. As autoras estudam as abordagens e concepções utilizadas por pessoas com treinamento específico em indexação de documentos para aplica-las aos modelos de IA. Nesse sentido, aumentam a importância da análise macrotextual para seleção dos índices. Como a base de dados utilizada para teste é um conjunto de artigos científicos em língua inglesa, as áreas do texto consideradas mais relevantes são o título, palavras chaves selecionadas pelo autor, resumo, citações, entre outras. Utilizando essas seleções como fonte de conhecimento semântico, as autoras propõem um *framework* de IA embasado em conceitos. São realizados diversos experimentos de indexação e submissão de pesquisas à base e os resultados demonstram que, entre outras conclusões, a indexação de conceitos é mais efetiva, em termos do índice F, do que a indexação por extração do texto completo.

Névéol, Rogozan e Darmoni (2006), por outro lado, reconhecem que a profusão de documentos disponíveis nas bases de dados online impossibilita a manutenção e atualização de índices manuais, sobretudo por causa do custo operacional de homem–hora. Assim, mecanismos de IA são requeridos. Utilizando uma base de dados de documentos da área médica em língua francesa, uma arquitetura de sistema é desenvolvida cujos resultados experimentais de pesquisas submetidas apontam índices de precisão comparáveis às bases de IM. A pesquisa

de Camara Junior (2007) também desenvolve um sistema de IA por sobre uma base de documentos da área jurídica em língua portuguesa cujos resultados de revocação e precisão são, em alguns casos, até superiores às bases de IM. Lahtinen (2000) igualmente dispõe de uma base indexada manualmente para comparação e treinamento, e propõe um modelo de IA híbrido que utiliza um *parser* sintático na vertente linguística e a contagem de frequência de termos na vertente estatística.

Concluindo a discussão, Zobel e Moffat (2006) apresentam um tutorial com técnicas chave de indexação. O ponto de vista é direcionado para a CC, com análise de aspectos mais técnicos do processo, desde armazenamento e construção de índices até avaliação de resultados de pesquisa. Além disso, um extenso levantamento bibliográfico da literatura sobre indexação de textos é exposto, servindo como boa referência para estudos e fundamentos na área. Já Pulgarín e Gil-Leiva (2004) apresentam uma análise bibliométrica da literatura de IA. Diversos aspectos são analisados, tais como distribuição de autores e trabalhos, obsolescência e dispersão. Um resultado bastante interessante apresenta-se na produção por tipo de documento. Uma linha crescente de quantidade de artigos científicos se desenha até o último quinquênio, quando há abrupta queda. Os autores atribuem a falta ou retardo de atualização de alguns BD. A quantidade de teses de doutorado, contudo, segue em constante crescimento. Conquanto o estudo já tenha mais de 10 (dez) anos, mesmo assim demonstra-se o quanto a IA ainda cativa constante interesse por parte da comunidade científica, mormente pela falta de consenso metodológico, o que deixa portas abertas para investigação. Esse é exatamente o ponto onde esta pesquisa deseja se inserir.

Concluindo o referencial teórico de IA, ele se associa a esta tese em sua aplicação pragmática. A IA é um dos vários empregos possíveis para PLN, e o levantamento procura demonstrar que essa é uma área de pesquisa que ainda demanda resultados consolidados. Além disso, os trabalhos apresentados permitem delinear o processo de IA e sustentar as decisões do projeto.

2.5. MARCO TEÓRICO

O marco teórico desta pesquisa é composto por extratos do referencial teórico ressaltando as abordagens diretamente adotadas. Quanto às ontologias, a definição de Gruber (1993) como a especificação explícita de uma conceitualização, e Guarino, Oberle e Staab (2009) como uma hierarquia de conceitos, são balizadores. Guarino (1998) identifica a ontologia construída nesta pesquisa como uma ontologia de aplicação, porque é desenvolvida a partir de documentos de uma área bastante restrita. Sua arquitetura, todavia, é de ontologia de alto nível, uma vez que o objetivo da mesma é modelar a significação de uma linguagem natural. Stevens e Lord (2009) ressaltam a análise de textos, indexação e vocabulário controlado como aplicações de ontologias, as quais são diretamente abraçadas nesta pesquisa. Hirst (2009), por fim, coloca que as ontologias podem prover interpretação para o sentido de palavras, que é precisamente a intenção deste trabalho.

Já em relação ao PLN, Jurafsky e Martin (2008) detalham todas as fases do processo sendo sua obra uma referência destacada na área, citada por praticamente todos os pesquisadores. Bird, Klein e Loper (2009) mantêm vivo o projeto do NLTK, que é a infraestrutura de *software* utilizada nesta pesquisa. Nirenburg e Raskin (2004) ensinam como realizar o processo com resultados de fases posteriores sendo retroalimentados nas fases anteriores para melhoria dos resultados. Essa é a abordagem utilizada no trabalho. Palmer (2010) identifica os desafios da fase de pré-processamento, enquanto Hippiisley (2010) ensina como realizar a análise morfológica. Para essa análise léxica é escolhida a utilização de métodos estatísticos utilizando abordagem de máxima entropia no MOM.

A análise sintática é teoricamente balizada pelas GLC de Chomsky (1956). Ljunglöf e Wirén (2010) descrevem os mecanismos e algoritmos clássicos de *parse*, e o algoritmo selecionado é o de Earley, o qual não exige gramáticas na FNC. Martins, Hasegawa e Nunes (2012) oferecem a GLC utilizada nesta pesquisa. A análise semântica, por fim, uma vez que o experimento não chega ao nível de profundidade pragmático, tem seu marco teórico em Goddard e Schalley (2010), Nirenburg e Raskin (2004) e Pustejovsky (1991). A abordagem selecionada para

esta pesquisa é a semântica léxica, onde o significado é atribuído para cada unidade lexical, e a combinação delas gera o sentido global. Nesse caso, as unidades lexicais assumem facetas diferentes em cada contexto, e isso precisa ser avaliado. Goddard e Schalley (2010) discutem os detalhes e implicações do processo. Pustejovsky (1991) introduz a abordagem de léxico gerativo para descrever as propriedades e relações entre conceitos. Nirenburg e Raskin (2004) detalham como devem ser estruturadas as fontes de conhecimento estático utilizadas para representação do significado textual. Assim, a ontologia construída lança mão desses pressupostos e suporta o desenvolvimento da análise semântica.

O marco teórico da pesquisa também é integrado pelo método de avaliação de sistemas de RI proposto por Savoy e Gaussier (2010), qual seja o índice F, que é a média harmônica entre os índices de revocação e precisão de pesquisas. Essa é a métrica utilizada para medição da efetividade da arquitetura de sistema proposta. Baeza–Yates e Ribeiro–Neto (2011) apresentam uma obra que é muito citada por pesquisadores da área, sendo também uma referência diferenciada para RI. Quanto à IA, a base se estabelece com Borko (1977), Lancaster (2004) e Robredo (2005) na explicação de como realizar o processo de indexação, destacando sua importância para RI. Anderson e Pérez–Carballo (2001), Moens (2000) e Souza (2006) discutem a IA propriamente dita e propõem metodologias para realização do processo. Esse último, em particular, utiliza uma abordagem híbrida com métodos linguísticos e estatísticos que inspira esta pesquisa.

Dois pressupostos deste trabalho, os quais emergem do levantamento teórico, são descritos. O primeiro apresenta que o prejuízo para a RI que a IA apresenta frente à IM não é significativo. A análise de custo e benefício indica um resultado muito favorável à IA, uma vez que o custo tende a 0 (zero). Já o segundo propõe que a análise semântica de PLN melhora a seleção de descritores para IA, incrementando, assim, a qualidade da RI. Esses pressupostos esboçam os trabalhos metodicamente executados permitindo uma clara compreensão do que efetivamente se deseja alcançar com a metodologia.

Concluindo, portanto, o referencial teórico procura situar este trabalho na pesquisa científica da área. Por meio da apresentação de outros estudos correlatos

ou próximos, relaciona-se esta investigação ao que se tem de resultados obtidos até então e propõe-se a extensão de produtos anteriores. Com isso, é possível demonstrar a originalidade desta pesquisa e utilizar trabalhos precedentes como marco teórico para suporte científico das propostas. Numa citação recorrentemente utilizada atribuída a Newton, só é possível enxergar mais longe apoiado sobre ombros de gigantes, e o referencial teórico objetiva exatamente a construção desse suporte.

3. METODOLOGIA

O conhecimento científico, como descrito por Marconi e Lakatos (2004), é factual porque lida com evidências de ocorrências ou fatos. Além disso, é contingente porquanto suas proposições são examinadas por meio de experimentação, e não exclusivamente pela argumentação. A experimentação é sistemática devida a sua ordenação lógica conectada com outras teorias e, principalmente, verificável por meio da replicação dos procedimentos e métodos. Por fim, o conhecimento científico não é absoluto, o que significa que teorias são amadurecidas ao longo do tempo para absorção de novas percepções cientificamente demonstradas.

Assim, a metodologia de pesquisa tem por alvo classificar e especificar os métodos utilizados para execução do trabalho. O objetivo é explicar as decisões tomadas e detalhar os procedimentos executados de forma a justificar os resultados alcançados e tornar o experimento replicável.

3.1. TIPO DE PESQUISA

Creswell (2009) afirma que há três tipos de desenho para pesquisas científicas, quais sejam quantitativo, qualitativo ou misto. Coloca ainda que as fronteiras entre as abordagens não são tão claras, e que uma pesquisa tende a ser mais quantitativa ou mais qualitativa. Esta pesquisa, em particular, apresenta-se com caráter mais qualitativo devido a suas características intrínsecas.

A pesquisa qualitativa é uma estratégia para aprofundar o estudo de uma teoria, ou hipóteses, por meio do exame de relacionamentos entre variáveis (CRESWELL, 2009). Nesta pesquisa, propõe-se que um módulo semântico de PLN pode oferecer índices de revocação e precisão considerados ótimos, ou seja, com valores calculados iguais ou próximos a 1 (um). As variáveis, portanto, são os descritores e os índices de resposta a pesquisas na base de dados utilizados para avaliação dos resultados. Tais variáveis são medidas por meio dos instrumentos de pesquisa de forma que os dados coletados possam ser sopesados. Marconi e

Lakatos (2004) diferenciam as pesquisas qualitativas das quantitativas precisamente pela forma de análise dos dados, sendo o uso da estatística adequado para essas. Este trabalho, por seu caráter mais qualitativo, não procura estender o universo de seus resultados por meio da análise estatística.

Perceba-se, por outro lado, que Marconi e Lakatos (2004) colocam que a pesquisa quantitativa exige amostras amplas e informações numéricas, enquanto os métodos qualitativos usam amostras reduzidas e análises psicossociais. Creswell (2009), no entanto, não concorda com a afirmação propondo que a pressuposição de que a pesquisa quantitativa trate de 'números' e a qualitativa trate de 'palavras' precisa ser amadurecida. Uma forma mais abrangente de reconhecer as diferenças entre elas pode ser a análise das suposições filosóficas da pesquisa, suas estratégias e métodos de aplicação. Esta pesquisa, em particular, apresenta um caráter mais qualitativo com amostra reduzida, porquanto a modelagem semântica de uma base muito grande não seria viável no período de realização do estudo.

Assim, Creswell (2009) propõe que a classificação de uma pesquisa deve ser realizada por três componentes inter-relacionados. O primeiro deles é a visão filosófica de mundo. Ela representa as pressuposições de alto nível, ou seja, mais conceituais, que o pesquisador traz para o estudo. Esta pesquisa tem forte caráter pós-positivista, no sentido de procurar as causas que geram os efeitos para os fenômenos. No caso, como realizar o processo de indexação (causa) para melhorar os resultados da RI (efeito). Além disso, o pós-positivismo preconiza o reducionismo, qual seja a redução do universo a variáveis que possam ser controladas e testadas por meio de experimentos. Por fim, acrescenta-se que essa visão filosófica propõe a observação empírica para coleta e medição de parâmetros com o objetivo de verificação de teoria. Isso significa que os dados, evidências e considerações racionais são utilizados objetivamente para aquisição de conhecimento. Este trabalho parte desses pressupostos para construção de seu plano de pesquisa e atingimento de seus resultados.

Ainda na visão filosófica de mundo, Creswell (2009) também explica que o pragmatismo é uma posição que, ao contrário do pós-positivismo, dá maior relevância às ações, situações e consequências frente às condições antecedentes.

Neste estudo, em particular, essas visões não se confrontam, mas se complementam no sentido de que o pragmatismo se preocupa com as aplicações práticas e soluções para os problemas. O foco se mantém no problema de pesquisa, e não nas causas ou no método. Nesse caso, o pesquisador tem certa liberdade para escolha de métodos, abordagens e suposições, assim como diferentes formas de coleção e análise dos dados. Esta pesquisa trata do problema da qualidade da RI, e algumas propostas são testadas para tratativa e melhoria dessa questão.

O segundo componente de classificação de uma pesquisa, na opinião de Creswell (2009) é a estratégia de pesquisa. Esta pesquisa, por apresentar uma vertente mais qualitativa, utiliza uma estratégia embasada em experimento. O experimento procurar desvendar se determinado tratamento ou abordagem influencia um efeito. Assim, o experimento deste trabalho se delinea com o controle dos mecanismos para IA de documentos e verificação de seus efeitos nos resultados de RI sobre a base.

Por fim, o terceiro componente representa o método de pesquisa propriamente dito, ou seja, quais formas de captura, análise e interpretação de dados serão realizadas. Esta pesquisa parte de instrumentos construídos pré-determinados para avaliação de dados de desempenho por meio de análise e interpretação. Assim, Creswell (2009) postularia a classificação deste trabalho como uma pesquisa com visão filosófica de mundo pós-positivista, associada à vertente pragmática, que utiliza estratégia de pesquisa experimental com instrumentação própria para avaliação de dados de performance sobre suporte não estatístico.

3.2. CARACTERIZAÇÃO DA AMOSTRA

Os documentos utilizados para avaliação dos resultados desta pesquisa consistem de laudos periciais de crimes cibernéticos produzidos pela perícia criminal da Polícia Federal (PF), o que define o universo da pesquisa. O laudo é um documento que tem por objetivo formalizar a autoria e materialidade de um crime após o vasto exame científico do corpo de delito deixado por uma infração penal. Nesse sentido, o laudo pericial muito se assemelha a um relatório acadêmico de

pesquisa científica. No laudo são minuciosamente consignados os exames realizados pelo perito e as respostas aos quesitos formulados (BRASIL, 2009). O Código de Processo Penal preconiza que o laudo seja elaborado por perito oficial, o qual, no atual ordenamento jurídico Brasileiro, para os crimes contra a União, é o Perito Criminal Federal (PCF) da PF.

O laudo é um documento com macroestrutura muito bem definida. Conquanto haja diversas áreas de exame pericial, tais como perícia de informática, perícia contábil, perícia de engenharia civil, perícia de meio ambiente, entre várias outras, e por extensão diversos tipos de laudos, ainda assim todo laudo apresenta um modelo padrão macroestrutural. Esse modelo se descreve pelos 6 (seis) elementos a seguir:

- Preâmbulo
- Descrição do material
- Objetivo
- Exame
- Conclusão
- Apêndice

Primeiramente, o laudo contém o preâmbulo. Nessa área é apresentada a motivação do laudo, ou seja, o evento que originou a realização da perícia. Assim como um juiz de direito, um PCF não pode agir de ofício, ou seja, um perito não pode deliberadamente realizar um exame pericial e elaborar um laudo. Ele deve ser instado a fazê-lo a partir da formulação de quesitos que devem ser respondidos por meio de exame científico. Qualquer entidade pode formular quesitos. A autoridade policial, na figura do delegado de polícia, é quem usualmente o faz, porém o membro do Ministério Público (MP), o juiz de direito ou até mesmo as partes do processo também podem quesitar. Assim o preâmbulo descreve os peritos que elaboraram o laudo, a autoridade que os designou para tanto, o solicitante, os documentos que acompanham o procedimento e, por fim, os quesitos propriamente ditos.

Já a segunda parte do laudo é a descrição dos materiais. Nessa seção detalham-se os vestígios que serão submetidos a exame. Essa é uma área onde é muito comum encontrarem-se fotos dos materiais para promover melhor descrição. Em crimes cibernéticos, os computadores ou mídias apreendidas são caracterizados nesse lugar. A terceira parte, por sua vez, apresenta o objetivo dos exames. De maneira geral, o objetivo de qualquer laudo é responder aos quesitos. Ocorre que a resposta ao quesito pode não ser um fim por si só, contudo um resultado consecutivo da busca de um objetivo. Isso deve ser precisamente descrito nesse seguimento, a fim de elucidar o leitor da finalidade dos exames.

A quarta parte, provavelmente a mais importante para IA, é a explicitação do próprio exame. Nessa área, o PCF pormenoriza o método utilizado para realização dos procedimentos e explica cada uma das fases do processo até o alcance de seus objetivos. Além disso, há o detalhamento das decisões tomadas durante a execução dos trabalhos e citação de referencial bibliográfico da área para fundamentação das mesmas. Essa é comumente a parte mais extensa de um laudo.

A última parte, por fim, é a conclusão. Nela são objetivamente expostas as conclusões a que se chegaram a partir da realização dos exames periciais. É de fundamental importância que as conclusões sejam cientificamente demonstradas a partir dos experimentos realizados, até para que os mesmos possam ser replicados, se necessário. Finalmente, na conclusão são respondidos os quesitos do preâmbulo. Um laudo ainda pode ter um conjunto de apêndices ou anexos acrescidos pelo perito caso o mesmo acredite que alguma informação adicional seja útil para compreensão do laudo. Nos laudos de crimes cibernéticos, por exemplo, é muito comum explicitar-se em apêndice a metodologia utilizada para garantir a integridade das eventuais mídias digitais que acompanhem o laudo, além de instruções para verificação. Com isso, o destinatário ao qual o laudo se dispõe pode atestar os conteúdos e averiguar a cadeia de custódia das provas periciais.

Percebe-se, destarte, que o laudo é um documento com uma estrutura muito bem definida, o que melhora os resultados da aplicação de PLN. Observe-se que o laudo não é um documento público. Enquanto no âmbito da PF, o laudo é sigiloso. Após a autoridade policial concluir o inquérito e enviá-lo ao MP, o laudo

ainda não tem publicidade. Após a análise do MP, caso o membro decida pelo oferecimento de denúncia, o laudo será enviado ao Poder Judiciário. A publicidade do laudo pericial somente ocorre quando o processo judicial estiver autuado no Tribunal de julgamento, e ainda assim apenas caso o processo não esteja em tramitação de segredo de justiça. A implicação desse fato nesta pesquisa é que em todo e qualquer procedimento realizado não é dada publicidade aos documentos utilizados. Além disso, a descrição das pesquisas, análises dos resultados e discussão das idiossincrasias são realizados de forma a nunca especificar qualquer investigação policial ou nome de indiciado.

Os laudos da PF encontram-se armazenados em um SI desenvolvido e mantido por peritos criminais da área de informática do órgão. O sistema, denominado Criminalística, tem por objetivo estruturar todas as tarefas administrativas afetas à perícia, tais como agendamento, gestão de pessoas, recepção de material, gestão eletrônica de documentos, entre várias outras. Além disso, o sistema procura organizar e representar o conhecimento produzido na perícia da PF por meio do fornecimento de repositório dos laudos periciais e ferramentas de pesquisa à base.

O problema, no entanto, consiste em que não há uma arquitetura da informação bem definida na construção do sistema. Isso significa que um documento só pode ser recuperado por meio de parâmetros armazenados de seus metadados, tais como nome dos peritos responsáveis pela elaboração, data do documento, assunto selecionado pelos autores, entre diversos outros. Além disso, uma indexação de todo o texto de cada laudo foi realizada para permitir a consulta por palavras chave. Esse tipo de indexação por extração, fartamente chamada *fulltext* na literatura, acarreta alguns problemas, os mais proeminentes quanto à perda de qualidade da RI, explicitada nos baixos índices de revocação e precisão de pesquisas. Isso representa um grande prejuízo para a perícia criminal da PF, e nisso consiste a contribuição pragmática deste estudo.

A amostra não probabilística extraída da base para o estudo são os laudos de crimes cibernéticos produzidos pela perícia criminal da PF, em todo o País, durante o primeiro trimestre do ano de 2012. A escolha do primeiro trimestre

se justifica pela distância temporal da produção dos laudos, o que aumenta a probabilidade dos documentos já estarem no âmbito do Poder Judiciário, onde sua publicidade aumenta. Essa seria uma questão resolvida em sua totalidade caso a amostra fosse selecionada a partir de uma considerável quantidade de tempo: o primeiro trimestre do ano de 2002, por exemplo. A aplicação da pesquisa na base mais atual é, entretanto, além de mais interessante do ponto de vista pragmático, mais útil para a PF. Assim, 2.285 (dois mil, duzentos e oitenta e cinco) documentos são recuperados do sistema Criminalística por meio dos seguintes parâmetros:

- Tipo de documento: Laudo
- Data de emissão: 01/01/2012 a 31/03/2012
- Unidade de registro: Todos
- Área do exame: Perícias de informática

Com essa amostra não probabilística, é realizada extração aleatória de 2/3 (dois terços) da base para treinamento do ferramental. Assim, 1.523 (um mil, quinhentos e vinte e três) laudos são selecionados para treinamento e 762 (setecentos e sessenta e dois) laudos são somados aos anteriores para avaliação. Conquanto a amostra seja não probabilística, extraída com o exclusivo objetivo de exemplificar a aplicação do ferramental desenvolvido a partir da arquitetura proposta, ainda assim sua variância pode ser analisada. O Quadro 1 apresenta a distribuição dos laudos de crimes cibernéticos da PF por unidade de registro, ou local de produção. A primeira coluna apresenta a unidade de registro. Já a segunda coluna mostra as quantidades de laudos de crimes cibernéticos percorrendo todo o espaço amostral da base de dados do sistema Criminalística até 31/12/2012, assim como suas respectivas porcentagens do total. A terceira coluna, por fim, explicita a distribuição dos laudos no período de extração da amostra, qual seja de 01/01/2012 a 31/03/2012, também com suas respectivas porcentagens do total.

Unidade de Registro	Total de Laudos		Laudos da Amostra	
Região Norte				
SETEC/SR/DPF/AC	755	1.370 %	43	1.881 %
SETEC/SR/DPF/AM	1.206	2.188 %	47	2.056 %
SETEC/SR/DPF/AP	528	0.958 %	5	0.218 %
SETEC/SR/DPF/PA	1.348	2.446 %	34	1.487 %
UTEC/MBA/DPF/PA	209	0.379 %	3	0.131 %

UTEC/SNM/DPF/PA	104	0.188 %	11	0.481 %
SETEC/SR/DPF/RO	1.317	2.389 %	76	3.326 %
UTEC/VLA/DPF/RO	93	0.168 %	0	0 %
SETEC/SR/DPF/RR	508	0.921 %	42	1.838 %
SETEC/SR/DPF/TO	583	1.057 %	4	0.175 %
Região Nordeste				
SETEC/SR/DPF/AL	749	1.359 %	27	1.181 %
SETEC/SR/DPF/BA	1.431	2.596 %	78	3.413 %
UTEC/JZO/DPF/BA	51	0.092%	0	0 %
SETEC/SR/DPF/CE	1.308	2.373 %	110	4.814 %
UTEC/JNE/DPF/CE	70	0.127 %	3	0.131 %
SETEC/SR/DPF/MA	437	0.793 %	14	0.612 %
UTEC/ITZ/DPF/MA	68	0.123 %	15	0.656 %
SETEC/SR/DPF/PB	906	1.644 %	35	1.531 %
SETEC/SR/DPF/PE	1.127	2.045 %	37	1.619 %
UTEC/SGO/DPF/PE	21	0.038 %	0	0 %
SETEC/SR/DPF/PI	489	0.887 %	23	1.006 %
SETEC/SR/DPF/RN	1.021	1.852 %	53	2.319 %
SETEC/SR/DPF/SE	330	0.598 %	41	1.794 %
Região Centro– Oeste				
INC/DITEC/DPF	7.093	12.871 %	191	8.358 %
SETEC/SR/DPF/DF	1.836	3.331 %	93	4.070 %
SETEC/SR/DPF/GO	1.716	3.114 %	38	1.663 %
SETEC/SR/DPF/MS	1.810	3.284 %	75	3.282 %
UTEC/DRS/DPF/MS	422	0.765 %	15	0.656 %
SETEC/SR/DPF/MT	2.023	3.671 %	28	1.225 %
UTEC/ROO/DPF/MT	66	0.119 %	10	0.437 %
UTEC/SIC/DPF/MT	57	0.103 %	2	0.087 %
Região Sudeste				
SETEC/SR/DPF/ES	1.240	2.250 %	63	2.757 %
SETEC/SR/DPF/MG	3.063	5.558 %	129	5.645 %
UTEC/JFA/DPF/MG	439	0.796 %	26	1.137 %
UTEC/UDI/DPF/MG	315	0.571 %	3	0.131 %
SETEC/SR/DPF/RJ	4.461	8.095 %	145	6.345 %
SETEC/SR/DPF/SP	4.906	8.903 %	203	8.884 %
NUTEC/CAS/DPF/SP	346	0.627 %	25	1.094 %
NUTEC/STS/DPF/SP	78	0.141 %	2	0.087 %
UTEC/ARU/DPF/SP	35	0.063 %	6	0.262 %
UTEC/MII /DPF/SP	407	0.738 %	24	1.050 %
UTEC/PDE/DPF/SP	226	0.410 %	18	0.787 %
UTEC/RPO/DPF/SP	884	1.604 %	39	1.706 %
UTEC/SJK/DPF/SP	265	0.480 %	71	3.107 %
UTEC/SOD/DPF/SP	72	0.130 %	14	0.612 %
Região Sul				
SETEC/SR/DPF/PR	3.387	6.146 %	142	6.214 %
NUTEC/FIG/DPF/PR	743	1.348 %	31	1.356 %
UTEC/GRA/DPF/PR	237	0.430 %	44	1.925 %
UTEC/LDA/DPF/PR	470	0.852 %	31	1.356 %
SETEC/SR/DPF/RS	2.018	3.662 %	66	2.888 %

UTEC/PFO/DPF/RS	51	0.092 %	2	0.087 %
UTEC/PTS/DPF/RS	100	0.181 %	8	0.350 %
UTEC/SMA/DPF/RS	215	0.390 %	13	0.568 %
SETEC/SR/DPF/SC	1.465	2.658 %	27	1.181 %
TOTAL	55.105	100 %	2.285	100%

Quadro 1 – Distribuição dos laudos de crimes cibernéticos da PF por unidade de registro.

As unidades de registro se compõem pelas Superintendências Regionais da PF nas capitais dos Estados da Federação, onde se localizam os Setores Técnico–Científicos (SETEC) da PF. Além desses, há Núcleos Técnico–Científicos (NUTEC) e Unidades Técnico–Científicas (UTEC) espalhadas em algumas cidades de maior porte. Por fim, o órgão central da perícia nacional, localizado em Brasília/DF, é o Instituto Nacional de Criminalística (INC) da Diretoria Técnico–Científica (DITEC) da PF. Ao total, há 54 (cinquenta e quatro) unidades de registro de laudos.

Percebe-se, pelo estudo estatístico do Quadro 1, que a distribuição da amostra segue, de maneira geral, o mesmo padrão da distribuição total da base. A média simples das diferenças das porcentagens entre a base completa e a base da amostra é de 0.637 % (zero ponto seiscentos e trinta e sete por cento). A média ponderada pela quantidade de documentos da base total é de 1.205 % (um ponto duzentos e cinco por cento), enquanto a média ponderada pela quantidade de laudos da base da amostra é de 1.053 % (um ponto zero cinquenta e três por cento). O desvio quadrático médio é de 0.142 % (zero ponto cento e quarenta e dois por cento).

Não obstante os resultados apresentarem uma aproximação expressiva, ainda assim é possível discutir pontos de melhoria. Primeiramente há que se considerar que, historicamente, as unidades de registro não foram criadas no mesmo momento. O panorama apresentado atualmente remete a Julho de 2009, quando foram instituídas todas as UTEC. Assim, a comparação da base da amostra com a base total é injusta, pois se comparam unidades com diferenças grandes de idades. Além disso, a perícia da PF tem adotado a política de descentralizar a produção de laudos, esvaziando a produtividade do INC. Ela ainda é significativa frente ao cenário nacional, porém claramente menor na amostra, 8.358 % (oito ponto

trezentos e cinquenta e oito por cento), do que na base total, 12.871 % (doze ponto oitocentos e setenta e um por cento), visto que a amostra é mais recente, onde essa política já se encontra mais institucionalizada.

Concluindo, percebe-se que a variância da amostra é muito pequena, o que, caso a pesquisa objetivasse a extensão dos resultados para o universo da base, aponta que a porção parece ser suficiente para representação do total. Como o objetivo do estudo é apresentar uma exemplificação da aplicação da arquitetura proposta, a reduzida amostra não probabilística foi selecionada de forma não aleatória e atende aos requisitos da tese.

3.3. INSTRUMENTO

Vários instrumentos são utilizados para a realização desta pesquisa. Alguns deles são desenvolvidos, outros adaptados, e alguns outros apenas utilizados como se apresentam. A composição de ferramental existente com o desenvolvimento de novos instrumentos é uma das inéditas contribuições deste trabalho. A descrição das ferramentas encontra-se realizada na mesma ordem em que as mesmas são utilizadas ou desenvolvidas, de forma a facilitar a leitura e compreensão do método adotado, além de ratificar a ordenação do procedimento.

O primeiro instrumento utilizado nesta pesquisa é o NLTK. Esse ferramental, programado e mantido por Bird, Klein e Loper (2009), possui um extenso conjunto de ferramentas e recursos para PLN. Por se tratar de um conjunto de artefatos genéricos, há esforço de programação para adaptação para língua portuguesa. O primeiro deles, por exemplo, é em relação à acentuação, o que embora em idioma inglês não seja relevante, para língua portuguesa é crucial. Além disso, os recursos nativos são para aplicação em idioma inglês, o que exige a construção de novos adaptadores ou recursos completamente originais. Isso remete desde a fase de pré-processamento de PLN, passando pela etiquetagem POS da análise morfológica até a construção da árvore de decomposição da análise sintática.

O NLTK encontra-se instalado em plataforma Unix com SO de código aberto Ubuntu versão 12.04 LTS com núcleo 3.2.0–31. As versões desse SO utilizam ano e mês de disponibilização para numeração, indicando que essa é de abril de 2012. Além disso, LTS é acrônimo para suporte de longo prazo, informando que essa versão do SO Ubuntu é suportada pela comunidade durante o prazo de 2 (dois) anos. Ademais, o NLTK é desenvolvido em linguagem de programação Python, o que exige a instalação desse ambiente de desenvolvimento. Implanta-se, assim, o Python versão 2.7.3 com atualização de 01 de agosto de 2012. O NLTK propriamente dito, por fim, está instalado na versão 2.0.3 de setembro de 2012. Todos os assessórios do NLTK são baixados e instalados localmente, mesmo aqueles que previamente já se sabia que não seriam utilizados, como as árvores sintáticas de treinamento em idioma alemão, por exemplo. Todos os *softwares* descritos são de código aberto, com licenciamento livre, não demandando qualquer aquisição de licença de uso ou pagamento de royalties, além de, evidentemente, não quebrar qualquer patente ou infringir direito autoral.

A instalação de aplicativos em ambiente Unix, infelizmente, não é tão trivial quanto se desejaria. As ferramentas de análise profunda de PLN do NLTK exigem a instalação do pacote NumPy, que é um pacote matemático de base para computação científica em linguagem Python. Trata-se de uma biblioteca de artefatos aritméticos complexos. Esse pacote precisa ser instalado, e exige como pré-requisito o pacote python-dev. Com essas duas instalações, que demandam minutos para serem concluídas, o ambiente está pronto para ser utilizado.

Em relação à análise morfológica, particularmente, descreve-se o segundo instrumento utilizado na pesquisa. Trata-se da floresta sintática, projeto de Freitas, Rocha e Bick (2008). A floresta é um *corpus* em idioma português morfossintaticamente anotado de forma semiautomática. A um extenso conjunto de documentos é aplicada ferramenta de *parse* automática a qual realiza a anotação das unidades lexicais sem supervisão. Após esse passo, esse conjunto passa por um processo de revisão chegando ao resultado final de 300.000 (trezentas mil) palavras revistas por linguistas e 3.800.000 (três milhões e oitocentas mil) palavras sem revisão. Esse projeto iniciou-se no ano 2000 e continua em plena evolução.

Esse recurso é utilizado para treinamento e avaliação de analisadores morfossintáticos.

Já para a análise sintática, a gramática do Curupira (MARTINS; HASEGAWA; NUNES, 2012) é empregada. Esse terceiro instrumento é incorporado aos reconhecedores disponíveis nativamente no NLTK para montagem da árvore de *parse*. Percebe-se, portanto, que a abordagem utilizada para realização das análises de PLN é híbrida. Utilizam-se métodos linguísticos para modelagem da linguagem e métodos estatísticos para atribuição e avaliação dos resultados.

Com o ferramental de PLN até o nível sintático estabelecido e testado, parte-se para a seleção dos documentos para teste do *framework*. Para realização desta tarefa utiliza-se o quarto instrumento desta pesquisa, o sistema Criminalística. Desse sistema é extraída a amostra discutida na Seção 3.2, qual seja um conjunto de laudos de perícias de crimes cibernéticos da PF. Não se realiza qualquer intervenção no aplicativo, apenas aproveita-se sua interface de pesquisa para extração manual dos laudos.

O quinto instrumento deste estudo, por sua vez, trata do módulo semântico de PLN. Uma ontologia para modelagem e extração de conhecimento é construída na forma preconizada por Nirenburg e Raskin (2004). Os autores definem que há quatro fontes de conhecimento estático para a análise semântica. As dependentes da linguagem são o conjunto de léxicos e o conjunto onomástico. A primeira estrutura contém um *corpus* anotado da linguagem. A segunda estrutura, uma lista de substantivos próprios, tais como nomes de pessoas, cidades, países, entre outros. Já as duas fontes de conhecimento estático independentes da linguagem são o repositório de fatos e a ontologia. O repositório de fatos contém registros de experiências passadas anotadas em um formato legível por computador. A ontologia é o instrumento desenvolvido nesta pesquisa, a qual é uma ontologia de aplicação com estrutura de ontologia de alto nível. A Figura 10 ilustra a relação entre essas quatro fontes de conhecimento estático.

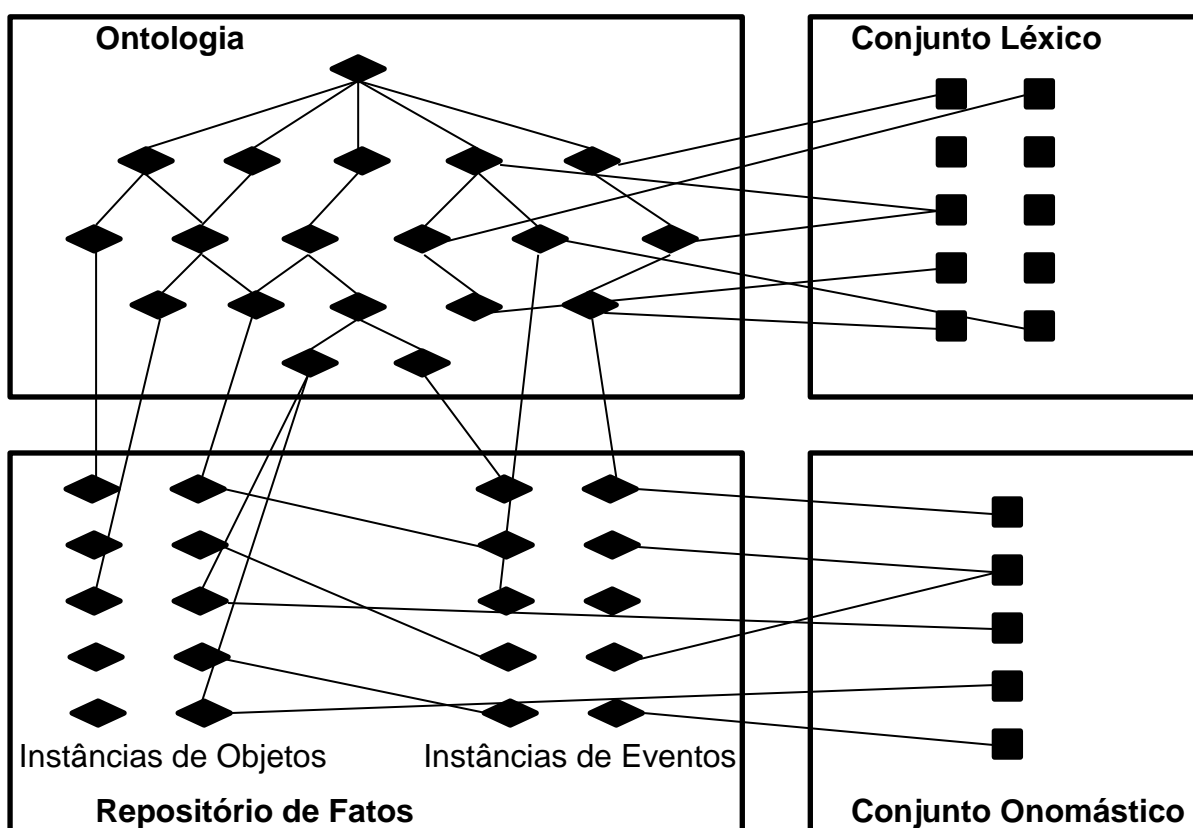


Figura 10 – Relacionamento entre fontes de conhecimento estático (NIRENBURG; RASKIN, 2004).

Utilizando-se o extrato de treinamento da amostra, formula-se a ontologia que é utilizada como fonte de conhecimento estático para o módulo de análise semântica de PLN. A ontologia proposta define um conjunto de categorias gerais aplicáveis para descrição da realidade linguística. Nirenburg e Raskin (2004) argumentam que algumas dessas categorias têm de ser representativas do senso comum, intenções, planos, ações, crenças, descrição de metac conhecimento, e, por fim, mecanismos de codificação de novas categorias geradas por aplicação de inferências do conhecimento já documentado. A ontologia, destarte, é constituída por um conjunto de conceitos, cada um deles representando nominalmente um conjunto de propriedades com valores especificados, pelo menos, parcialmente. Além disso, um conjunto de relações entre esses conceitos. Essa ontologia segue o seguinte modelo, a partir dos pressupostos da abordagem de léxico gerativo de Pustejovsky (1991):

- Conceito: conjunto de propriedades
 - Definição: acepção do conceito em linguagem natural

- Agente: entidade que causa ou é responsável por uma ação
- Tema: entidade manipulada por uma ação
- Paciente: entidade afetada por uma ação
- Instrumento: objeto ou evento utilizado para executar uma ação
- Fonte: ponto de partida de uma ação
- Destino: ponto de chegada de uma ação
- Lugar: localização onde um evento acontece
- Rota: a rota por onde uma entidade viaja
- Meio: estilo por meio do qual alguma coisa é realizada
- Índice: indicativo booleano para utilização como descritor no processo de IA independente de quaisquer outras análises
- Relação
 - É um: relação de herança entre conceitos
 - Sinônimo: relação de sinonímia entre conceitos
 - Hiperônimo: relação de hiponímia entre conceitos
 - Holônimo: relação de meronímia entre conceitos
 - Valor: mensuração para a relação
 - Sem: restrição para a relação
 - Padrão: relação default entre os conceitos
 - Relaxável a: extensão aceitável para violação da restrição da relação
 - Não: extensão não aceitável para violação da restrição da relação
 - Medida padrão: unidade de medida para mensuração
 - Inv: indicação de que a relação é inversa de outra
 - Espaço temporal: fronteira temporal onde determinado fato foi verdade
 - Origem: origem do elemento informacional que foi utilizado para construir a relação

Seguem 2 (dois) exemplos de entradas da ontologia. O Exemplo 1 é importado de Nirenburg e Raskin (2004), enquanto o Exemplo 2 é extraído da

ontologia desta pesquisa. Em letras maiúsculas os elementos da ontologia. Entre colchetes, a explicitação da relação.

Exemplo 1:

- PAGAMENTO
 - Definição
 - Compensar alguém por produtos ou serviços prestados
 - Agente
 - [sem] HUMANO
 - [relaxável a] ORGANIZAÇÃO
 - Tema
 - [padrão] DINHEIRO
 - [sem] FORMA DE PAGAMENTO
 - [relaxável a] EVENTO
 - Paciente
 - [sem] HUMANO
 - [relaxável a] ORGANIZAÇÃO

Exemplo 2:

- CRIME
 - Definição
 - Fato típico, ilícito e imputável descrito pelo Código penal (CP)
 - Agente
 - [sem] HUMANO
 - [não] ORGANIZAÇÃO
 - Paciente
 - [sem] HUMANO
 - [relaxável a] ORGANIZAÇÃO
 - Instrumento
 - [sem] ARMA
 - [sem] COMPUTADOR
 - Meio

- [padrão] MATERIALIDADE
- Índice
 - *TRUE*

3.4. PROCEDIMENTO

O primeiro procedimento desta pesquisa é sistematizar o funcionamento do ferramental de análise morfosintática de língua natural portuguesa. Para tanto, é utilizado o NLTK, como motor, a floresta sintática, como *corpus* anotado, e a GLC do Curupira. O NLTK é escrito em linguagem de programação Python, o que determina o aprendizado da codificação para realização das customizações necessárias. As ferramentas nativas do NLTK têm de ser treinadas com o *corpus* para obtenção de resultados satisfatórios nas análises. O etiquetador POS e o decompositor, utilizados nas análises morfológica e sintática respectivamente, embasam seus resultados no treinamento realizado a partir da floresta sintática e da incorporação da GLC do Curupira.

Após a instalação da infraestrutura de análise, parte-se para a seleção dos documentos. A amostragem é realizada como descrito na Seção 3.2 desta tese, e os laudos são armazenados para processamento. O procedimento de recuperação dos documentos utiliza as interfaces padrão de consulta do sistema Criminalística. Já a extração dos documentos é estritamente manual, uma vez que não é autorizado realizar qualquer intervenção no código fonte do sistema nem tampouco acesso direto ao BD. Os mesmos são armazenados em repositório e numerados de acordo com sua respectiva data de emissão, em formato invertido (AAAA-MM-DD), associado a um sequencial numérico de 4 (quatro) dígitos iniciando em 0001 (um). Desta forma, o primeiro documento, por exemplo, é o '2012-01-02 0001.txt', o ducentésimo documento é o '2012-01-11 0200.txt', e assim sucessivamente.

Com os documentos selecionados, constrói-se a ontologia para a análise semântica. Utilizando o segmento da amostra para treinamento, uma ontologia é desenvolvida para suporte de significado. Esse esquema de representação do conhecimento evolui durante toda a realização da pesquisa, em um processo de

retroalimentação constante como preconizado na engenharia de ontologias. A importância da ontologia no processo de IA consiste no estabelecimento das relações entre os conceitos.

Assim, inicia-se a realização do PLN propriamente dito, seguindo suas fases características. A primeira fase, o pré-processamento, versa pela normalização dos documentos, extração das unidades lexicais e delimitação das sentenças. Os documentos extraídos do sistema Criminalística encontram-se nos mais diversos formatos. Há laudos em formato Microsoft Office Word até a versão 2003 (doc) e laudos em formato Microsoft Office Word versão posterior a 2007 (docx). Além desses, há vários documentos em formato Open Office (odt) e *Portable Document Format* (pdf). Alguns documentos se encontram em estado texto puro (txt) e um documento, em particular, está em formato Microsoft Office Excell versão 2010 (xlsx). Um esforço considerável é realizado, destarte, para padronizar e normalizar os documentos em um formato legível pelo NLTK. Essa tarefa foi realizada manualmente concomitantemente à extração de cada documento do sistema Criminalística. O formato uniforme escolhido para gravação dos laudos é, por motivos de facilidade de manipulação e economia de espaço, o texto puro (txt). A escolha se dá, além dos motivos citados, pela não necessidade de manutenção de formatação textual original.

Após a normalização do *corpus*, as ferramentas nativas do NLTK para tokenização e segmentação de orações são utilizadas. O treinamento com base anotada em língua portuguesa é fundamental para alcance de resultados. Um *script* em linguagem de programação Python é escrito para acionar as bibliotecas do NLTK e realizar a tokenização dos textos. Assim, cada unidade lexical reconhecida é armazenada em uma estrutura de dados de rápida navegação permitindo a realização de várias tarefas. A mais importante delas é a contabilização estatística de ocorrências, que é um importante parâmetro para decisão de seleção de descritor no processo de IA. O NLTK utiliza o caractere espaço em branco para delimitação de *tokens*. A heurística associada à utilização do dicionário do *corpus* permite reconhecer se algum outro caractere deve ser utilizado como separador, caso a caso. Uma palavra composta separada por um hífen, por exemplo, é reconhecida como uma única palavra se estiver contida no dicionário; ou, caso contrário, é

dividida em duas unidades, com o hífen sendo o caractere separador. Ressalte-se que não há lematização no processo de tokenização, ou seja, a estrutura de dados de armazenamento não recebe os radicais e as regras de formação de cada unidade lexical. Isso prejudica o sistema no tocante ao espaço necessário para repositório de processamento. No contexto desta pesquisa, em particular, devido à reduzida amostragem, o problema não se apresentou significativamente.

Uma importante observação procedimental se faz na fase de pré-processamento. Percebe-se que vários documentos em diferentes formatos são normalizados para um padrão único, legível. O pré-processamento responsabiliza-se precisamente por isso, comportando-se como o primeiro filtro por onde dados não estruturados são submetidos para iniciar o processo de organização. Assim, caso os dados sejam ainda menos estruturados, por exemplo, com a utilização de imagens, gravações de áudio ou vídeos, a fase de pré-processamento é o analisador que prepara todos os documentos e extrai os insumos para as fases seguintes, quais sejam os *tokens* e sentenças devidamente armazenados. Nesse aspecto, a escolha do tipo de documento só é relevante até o pré-processamento. A partir desta etapa, o tratamento é igual para qualquer informação recebida, o que é um resultado importante para a solução dos problemas de dependência discutidos anteriormente em PLN. Isso significa que uma arquitetura de PLN treinada para determinado *corpus* pode ser utilizada em outro conjunto de documentos de formato divergente alterando-se, em princípio, apenas o conjunto de ferramentas e procedimentos da fase de pré-processamento. Esse resultado, para a extensão das conclusões desta pesquisa, é muito relevante.

Com os *tokens* e orações devidamente delimitados, parte-se para a segunda fase de PLN, qual seja a análise morfológica. Ao mesmo *script* desenvolvido anteriormente para o pré-processamento são adicionadas as chamadas a procedimentos para a análise léxica. O resultado é o acréscimo de uma nova coluna à estrutura de dados que contém todos os *tokens* extraídos do texto para armazenamento de sua etiqueta léxica. O NLTK utiliza uma abordagem de máxima entropia para realização da etiquetagem, utilizando a floresta sintática como *corpus* de treinamento.

A partir deste resultado, a terceira fase de PLN se inicia. Na análise sintática, um dos *parsers* nativos do NLTK, ao qual se pode associar o algoritmo de Earley, é utilizado em conjunto com a GLC do Curupira. As sentenças segmentadas no pré-processamento são enviadas ao sistema dedutivo com as respectivas etiquetas POS de cada unidade lexical. O resultado é um conjunto de árvores de derivação de cada oração. A avaliação é feita também com abordagem de máxima entropia utilizando a anotação da floresta sintática.

O próximo procedimento interrompe o PLN e inicia a IA. Para indexação, o caráter híbrido desta pesquisa se ressaltava. Na vertente linguística, utilizam-se os sintagmas nominais selecionados a partir das árvores de derivação das sentenças. Como discutido na Seção 2.2, Kuramoto (1999), Souza (2006), Chaudiron (2007) e Maia (2008) usam sintagmas nominais para IA e classificação argumentando que essa é melhor estrutura léxica que contém significado com qualidade para indexação, comparando-se a escolha de palavras isoladas. Na vertente estatística, a razão da frequência do termo no documento com a frequência inversa na base (TF-IDF), como descrita por Baeza-Yates e Ribeiro-Neto (2011), é utilizada para seleção de descritores. Esses mecanismos são associados para obtenção de melhoria de resultados. Ao script escrito em linguagem Python para acionamento das primitivas do NLTK são acrescentadas as instruções para IA nesses pressupostos.

Com os candidatos a descritores selecionados, o PLN retorna ao foco com a análise semântica. A ontologia construída é utilizada para reconhecimento e validação dos descritores, além de proposta de novos índices. Os sinônimos, holônimos e hiperônimos são automaticamente selecionados. Ressalte-se que a ontologia não analisa exclusivamente os descritores pré-selecionados, no entanto todas as unidades lexicais e sintagmas passam pelo crivo ontológico. Isso se justifica porque os conceitos ontologicamente anotados com a propriedade 'Índice' são selecionados como descritor independente das análises anteriores, porquanto sua importância latente no texto. O crime cuja materialidade é descrito em um laudo, por exemplo, deve fazer parte do índice independente de qualquer análise linguística ou estatística realizada.

O último procedimento, por fim, consiste na exemplificação da arquitetura proposta por meio da discussão dos resultados da pesquisa. A amostra de documentos é indexada automaticamente e parâmetros textuais de pesquisa são avaliados frente aos descritores selecionados. Os resultados são calculados utilizando o índice F. Não há comparação aos resultados de buscas textuais no sistema Criminalística porque, como explicitado anteriormente, o Criminalística não tem IM em sua base de dados. A única atribuição de descritores aos documentos é a indexação por extração em todo o texto, o que oferece resultados irrelevantes para buscas textuais. Dessa forma, uma comparação com tal arquitetura não traz parâmetro razoável de conferição ou agrega sugestões de melhorias, além de não ser cientificamente extensível a comparação do experimento em uma amostra não probabilística pequena com o universo. O objetivo é exclusivamente a exemplificação e análise da proposta de arquitetura e funcionamento das implementações.

Concluindo, a Figura 11 ilustra e sintetiza os procedimentos de pesquisa para construção do método de IA proposto.

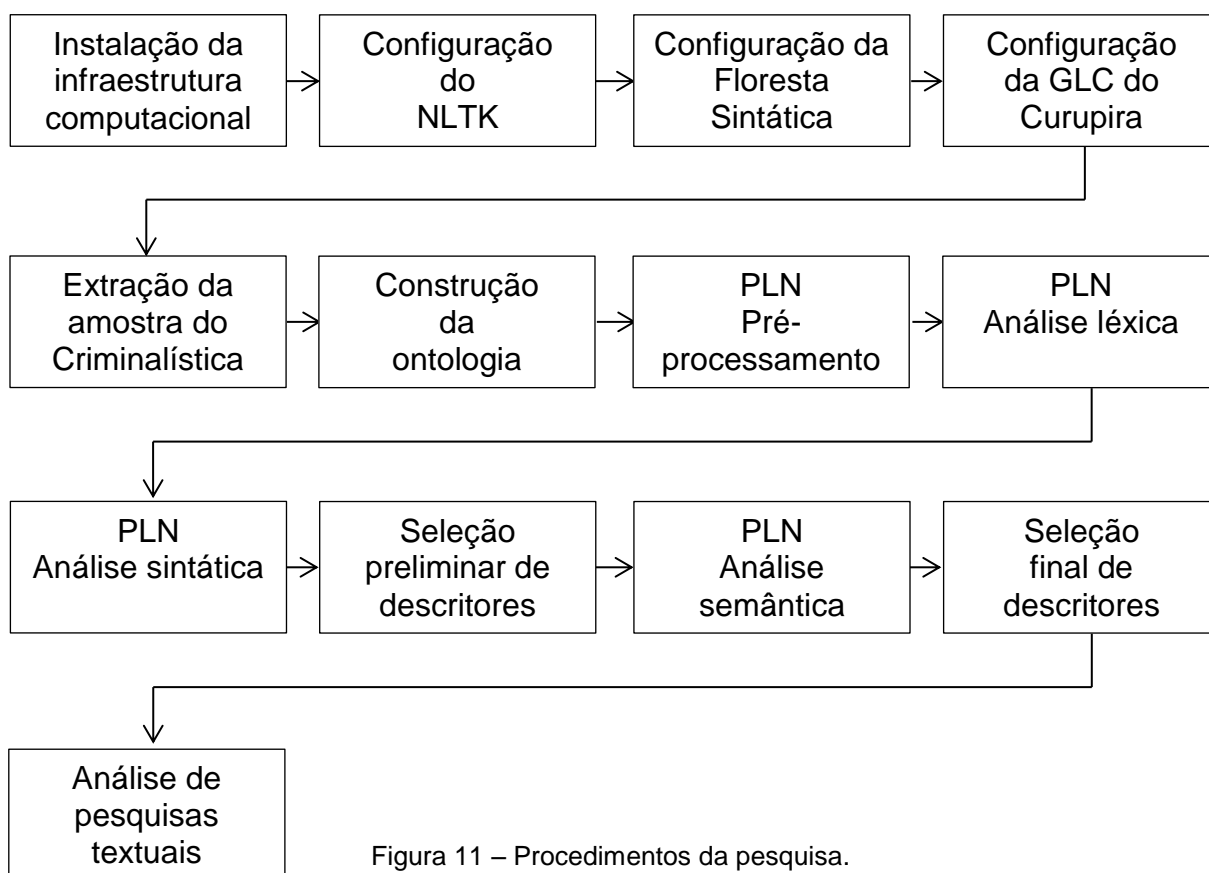


Figura 11 – Procedimentos da pesquisa.

4. PROPOSTA DE ARQUITETURA

Este capítulo objetiva explicitar detalhadamente a proposta de arquitetura de SI para IA de documentos em língua portuguesa utilizando PLN em nível semântico. Para tanto, são descritos os artefatos produzidos e os procedimentos executados. Além disso, é analisada a aplicação do ferramental desenvolvido com o objetivo de avaliar os resultados em uma pequena amostra, exemplificando a utilização.

4.1. DESCRIÇÃO DA ARQUITETURA

Primeiramente há que se comentar a construção da ontologia, que é o centro da análise semântica desta pesquisa. Guarino (1997) explica que ao se desenvolver uma ontologia, deve se incluir uma estrutura, não exclusivamente taxonômica, e que todas as suas relações devem existir em termos de seus significados. Essa orientação é seguida durante todo o processo de organização, de forma a permitir a extração do significado textual a partir da utilização da ontologia na análise semântica de PLN.

Para a estrutura taxonômica da ontologia, utiliza-se um tesouro. O tesouro escolhido é o tesouro jurídico do Superior Tribunal de Justiça (STJ), já utilizado por Camara Junior (2007) como esquema de representação do conhecimento para suporte à IA. O universo dos documentos utilizados na pesquisa, laudos de crimes cibernéticos da perícia criminal da PF, trata de matéria penal, coberta pelo tesouro jurídico. As relações ‘termo geral’, ‘termo específico’ e ‘termo relacionado’, no tesouro, são consideradas equivalentes às relações ontológicas ‘é um’, ‘hiperônimo’ ou ‘holônimo’, avaliando-se cada caso para descoberta da mais específica afinidade. A relação ‘inv’ é utilizada para inverter a associação entre os conceitos. Já a relação ‘use’ do tesouro é tratada para a relação ‘sinônimo’ da ontologia.

Com a estrutura taxonômica estabelecida, a proposta de arquitetura prescreve o treinamento do ferramental a partir da amostra do *corpus*, com objetivo de adição na ontologia. Nesta tese, cada um dos documentos é analisado

individualmente, em processo estritamente manual, e cada unidade lexical ou sintagma é extraído e acrescido na ontologia compondo uma relação ontológica de significação. As diversas facetas que um conceito adquire no contexto, que só é possível ser avaliado por meio da análise pragmática, são descobertas e descritas na ontologia, de forma que seja possível que o processo automático de análise semântica de PLN possa se beneficiar da representação formal.

O processo de construção manual da ontologia requer esforço considerável, pois envolve a extração manual das unidades lexicais dos documentos, assim como suas respectivas classificações e decisões para posicionamento ontológico. Por esse motivo não é possível utilizar, no contexto deste estudo, uma amostra probabilística de documentos, uma vez que seria exigida uma base muito ampla. O aprendizado automático de ontologias, outrossim, tem sido objeto de outras pesquisas. Uma delas encontra-se no mesmo Programa de Pós-Graduação em que esta tese se insere, qual seja o projeto já qualificado de José Marcelo Schiessl titulado Construção Automática de Axiomas no Contexto das Ontologias. Acredita-se que esta etapa da arquitetura, no futuro, possa incorporar o tratamento automático do *corpus* com o objetivo de auxiliar a construção da ontologia.

Em relação ao PLN propriamente dito, a primeira questão a se discutir é o pré-processamento textual. Como altercado na Seção 2.2.1, essa fase é responsável pela triagem documental e pela padronização e tokenização dos textos para construção do *corpus*. As estratégias de amostragem apontadas na Seção 3.2 demonstram como é realizada a triagem documental. Além disso, a arquitetura proposta demanda que os documentos sejam entregues em formato texto puro, sem formatação. Assim, nessa fase os documentos devem ser convertidos de quaisquer formatos em que se apresentem originalmente para arquivos em texto puro (txt).

Nesta pesquisa, essa etapa é realizada manualmente, convertendo-se cada documento extraído do sistema Criminalística para sua respectiva forma canônica, sem formatação. Essa decisão é tomada devido a que a avaliação proposta para o estudo determina o conhecimento de todos os documentos da base, para cálculo da revocação. Por conseguinte, à medida que os documentos recebem

tratamento de transformação, seu conteúdo é estudado e catalogado para julgamento dos resultados de submissão de pesquisas. Nada impede, contudo, que em qualquer aplicação desta proposta de arquitetura esse primeiro passo de pré-processamento seja automatizado por meio de um aplicativo conversor de formatos. Tal programa pode ser facilmente desenvolvido ou podem ser utilizados *softwares* livres disponíveis para execução da tarefa.

Uma vez que todos os documentos estejam padronizados, a próxima etapa do pré-processamento é a tokenização do texto. O objetivo é determinar todas as unidades lexicais e sentenças de cada um dos documentos. Para isso, o módulo de tokenização do NLTK é acionado via script em linguagem de programação Python. Por meio do aplicativo 'PlaintextCorpusReader', um conjunto de listas é alimentado a partir da leitura de cada um dos arquivos da amostra. Ao final dessa etapa, é possível selecionar individualmente cada um dos membros do conjunto de palavras e orações dos textos.

Concluído o pré-processamento, a arquitetura proposta preconiza que a análise morfológica seja realizada no *corpus*. Cada unidade lexical de cada um dos documentos decompostos é estudada de forma a identificar sua classe gramatical. O módulo léxico do NLTK é empregado para tanto, utilizando uma abordagem estatística de máxima entropia no MOM. O *corpus* da floresta sintática fornece o treinamento necessário para que o motor estatístico possa produzir suas inferências e classificar os *tokens*. A taxa de acerto nessa fase do processamento linguístico corrobora o emprego de estratégias de máxima entropia em diversas pesquisas de PLN e, sobretudo, a qualidade do *corpus* de treinamento. Em alguns laudos, por exemplo 2 (dois), 20 (vinte), 25 (vinte e cinco), 73 (setenta e três), entre outros, a taxa de acerto de classificação morfológica é de 100% (cem por cento).

O conjunto de etiquetas utilizadas na classificação POS recomendado pela arquitetura proposta nesta pesquisa encontra-se no Quadro 2.

Etiqueta	Classe Morfológica
ADJ	Adjetivo
ADV	Advérbio
CNJ	Conjunção
DET	Artigo
FW	Palavra estrangeira
N	Substantivo
NP	Substantivo próprio
NUM	Numeral
PRO	Pronome
P	Preposição
UH	Interjeição
V	Verbo

Quadro 2 – Etiquetas POS para a classificação morfológica.

Embora a análise léxica tenha apresentado um bom resultado, há que se comentar um aspecto de grande dificuldade nessa fase. Um exame de crime cibernético bastante comum tem se apresentado na perscrutação de equipamentos de telefones celulares e *smartphones*. Com a evolução tecnológica desse tipo de aparelho, cada vez mais informações podem ser armazenadas, com significativo valor forense. O conjunto de ligações realizadas por um telefone, ou mensagens enviadas por meio de serviço de mensagens curtas (SMS), pode confirmar a construção de uma rede de contatos. A agenda de contatos e compromissos também contribui para tal mapeamento. A capacidade de armazenamento interno de aparelhos, até os relativamente simples, permite a guarda de um grande conjunto de documentos. Alguns equipamentos ainda são capazes de ler e gravar dados em dispositivos de memória externa, compatíveis com computadores fixos e portáteis, o que alavanca a utilização desse tipo de aparelho como *backup* de informações. Percebe-se, por meio desses exemplos, o quanto a quantidade de perícias em telefones celulares e *smartphones* está crescendo atualmente, mormente quanto ao barateamento e popularização desses equipamentos.

Os laudos de aparelhos celulares e *smartphones* são documentos triviais, contendo a extração de agenda telefônica, ligações discadas e recebidas, mensagens SMS enviadas e recebidas, mensagens de correio eletrônico enviadas e recebidas, entre outras informações relevantes para a investigação. O desafio deste tipo de exame não é a produção do laudo propriamente dita, mas sim a extração das informações do equipamento, as quais podem estar apagadas, criptografadas,

armazenadas em formato não padronizado, ou maliciosamente escondidas. Ocorre, entretanto, que para mensagens SMS, em particular, uma nova espécie de linguagem se desenvolve e se torna padrão para esse veículo de comunicação. Essa linguagem possui diversas simplificações e erros gramaticais, que, em princípio, têm por objetivo facilitar e acelerar a produção do conteúdo. O Quadro 3 apresenta alguns exemplos de laudos da amostra com extratos de mensagens SMS trocadas entre remetentes e destinatários.

Número do Laudo	Mensagem SMS
21	- Eu ã vou ficar te ligando pra falar a hora pq vc tem relógio - Atendi é melho pra vc amanhã não adinta vou faze vc perde sei inpreguinho vc não chama seu pp
48	- ué aceito pq naum rsrs - mais dai vc manda o dinheiro e tals ? - aff meu , vctá se achando - coloka ae - vc ta +- mas axo que vai levar jeito né
134	- depois c reclama q eu n te ligo - fike de olho na sua namoradilha quando ela sai da facul
140	- Eu posso colok to indo na rua agora. Mas colok ai tbm pra agente se fala por msgn tbm
142	- Nossa linda pensa num cara kebrado sou eu to so o po axo ke vou xegar umas dez
271	- da p vc vim 8*h.q eu vo sair cedo.a quanto vc t entregando aki?
317	- Eu to ak Vc vai vim me burcar
548	- Vms entaum fazer um hh hj, espero todos lah, gde bjo
1.373	- dexa pro fim de tarde pode ser.abraso
1.384	- So to esperando minha irma chega p mim i ai te passa os cheque.ta?

Quadro 3 – Laudos de aparelhos celulares com mensagens SMS.

É de se notar, destarte, que esse tipo de construção linguística é difícil de ser interpretada automaticamente. O motor de PLN se ilude com os erros gramaticais e o não reconhecimento de unidades lexicais do idioma, o que prejudica a qualidade da análise. Nesse contexto, a análise estatística oferece vantagens frente à análise estritamente linguística. A extensa ocorrência de unidades lexicais como ‘vc’ substituindo ‘você’, ou ‘pq’ substituindo ‘porque’, por exemplo, levam o motor de máxima entropia a reconhecer essas construções e, em alguns casos, classificá-las corretamente. A análise linguística tem dificuldade em reconhecer e

tratar erros léxicos e gramaticais, o que inviabiliza a qualidade do exame neste tipo de redação, a qual mais se aproxima da linguagem falada do que da escrita.

Nesta pesquisa, não se cogita acrescentar ao *corpus* da floresta sintática documentos anotados com esse tipo de linguagem. Um conjunto de textos com entradas do tipo 'vc' anotadas como pronome poderia auxiliar o engenho estatístico a reconhecer essas construções. Não é realizado, no entanto, tal experimento nem identificadas vantagens ou prejuízos que tal abordagem pode trazer para a qualidade da análise morfológica.

Concluída a análise léxica, o próximo passo de PLN é a análise sintática. Na abordagem híbrida proposta por esta pesquisa, essa fase é realizada por estratégia linguística. As ferramentas de análise morfológica do NLTK permitem que se agrupem as unidades lexicais por uma distância fixa. Com isso, é possível construir bigramas, trigramas ou n-gramas, quais sejam conjuntos de duas, três ou 'n' unidades lexicais próximas, respectivamente. Esse tipo de abordagem, muito utilizada para IA, tem custo computacional baixo, e traz vantagens frente à indexação por palavras únicas, porém não é tão elaborado quanto a realização da análise sintática para extração dos sintagmas nominais. Nesta pesquisa, a árvore de *parse* sintática é construída para cada sentença de cada um dos documentos e os sintagmas nominais extraídos como candidatos a descritores dos laudos.

A primeira tarefa a ser executada, para atingir esse objetivo, é realizar o carregamento da gramática do Curupira para o reconhecedor do NLTK. Algumas alterações têm de ser realizadas no formato em que a gramática está escrita para que a mesma seja reconhecida. Primeiramente, a gramática do curupira está codificada em um conjunto de quadros, os quais têm de ser reescritos como regras inteligíveis para o NLTK. As regras precisam ser registradas no seguinte formato:

Símbolo não terminal -> Símbolo não terminal | 'Símbolo terminal' | ...

Onde o ícone '->' representa uma regra de produção, ou seja, demonstra que um determinado símbolo pode ser decomposto e um ou mais símbolos, terminais ou não terminais. Além desse, o caractere '|' representa o operador lógico

‘OU’, o que significa que um símbolo pode ter várias decomposições válidas reconhecidas pela gramática. Por fim, caso a regra chegue a um símbolo terminal, o mesmo deve ser envolvido por aspas simples. O processo de converter as tabelas da gramática do Curupira em regras no formato reconhecido pelo NLTK é manual, e só precisa ser realizado uma única vez, dado que apenas se a gramática sofrer alterações em suas regras é necessário que a mesma seja atualizada no NLTK.

O *parser* selecionado para realização da construção da árvore sintática é o algoritmo de Earley. Uma implementação do mesmo é escrita em linguagem de programação Python utilizando o ‘RecursiveDescentParser’ do NLTK. Esse é um reconhecedor *top-down* que parte dos símbolos não terminais até chegar recursivamente aos símbolos terminais nas folhas da árvore. Conquanto o processo de análise sintática não apresente uma acurácia tão significativa quanto a da análise morfológica, ainda assim os resultados permitem a extração dos sintagmas nominais que são utilizados como candidatos a índices dos documentos.

Por fim, a análise semântica completa o trabalho de PLN na proposta de arquitetura de IA. Nela, a ontologia construída é utilizada como esquema de representação do conhecimento para extração de informação semântica do texto. O objetivo final desta pesquisa é IA, ou seja, não é desenvolvida representação de significado textual, tais como axiomas em lógica de primeira ordem ou identificação de primos semânticos da metalinguagem semântica natural. Nenhum pressuposto da semântica composicional é utilizado, sendo a semântica léxica mais adequada para o objetivo de seleção automática de descritores.

Observe-se que a qualidade dos documentos selecionados para o experimento beneficiam as análises de PLN, de maneira geral. Um laudo pericial é um documento com características peculiares, o qual prescinde de figuras de linguagens, metáforas, estilística. É um texto objetivo, direto, que preza pela linguagem o mais clara possível, gramaticalmente impecável. Assim, os motores de análise não se iludem com erros de linguagem, ou hipérbatos desnecessários, que potencialmente dificultam a resolução de anáforas, melhorando, por extensão, a qualidade do PLN.

A arquitetura proposta orienta, por conseguinte, o processo de IA. A primeira fase parte da identificação das unidades lexicais, da análise morfológica, e dos sintagmas nominais, da análise sintática, para a contabilização da frequência de ocorrência nos documentos. Em outras palavras, na primeira fase a análise léxica é empregada para identificação das unidades lexicais e a análise sintática para extração dos sintagmas nominais. Com isso, aplica-se o cálculo descrito na Seção 2.4 para elencar os candidatos preliminares a descritores, ponderando-se a quantidade de ocorrências de um termo em um documento com sua ocorrência na base (TF-IDF). Esse processo é realizado utilizando-se as primitivas de contagem de frequência do NLTK.

Já no segundo passo, a ontologia construída é utilizada para seleção definitiva dos descritores. Cada um dos índices pré-selecionados é analisado sob o foco da ontologia para identificação de suas relações ontológicas. O objetivo é que a indexação realizada não seja por termos, todavia por conceitos. Assim, o processo de IA seleciona termos sinônimos, holônimos e hiperônimos para composição dos descritores.

Além disso, o texto original também é analisado sob o crivo da ontologia para verificar se há informação relevante para indexação que tenha passado despercebida na análise estatística de frequência. A ontologia possui uma propriedade para o conceito denominada 'Índice' a qual indica se o mesmo deve ser selecionado como descritor independente de quaisquer apreciações anteriores. Todas as previsões legais do CP para crimes cibernéticos estão com essa anotação ontológica, por exemplo. Caso o laudo apresente um termo reconhecido pela ontologia como descritor, o mesmo é selecionado se ainda não o tiver sido. Isso corrobora o caráter híbrido da arquitetura proposta, porquanto a análise estatística é empregada para seleção preliminar dos descritores, utilizando critérios estritamente matemáticos, enquanto a análise linguística parte de conhecimento semântico-ontológico para ajuste e incremento da seleção inicial.

Nesse aspecto, é relevante considerar que nomes de indiciados ou parceiros que sejam agentes da execução de uma atividade criminosa que deixou corpo de delito para ser periciado são importantes descritores de um laudo pericial.

Ocorre, contudo, que o nome do indiciado não é algo que se repete ao longo do texto de um laudo de forma que a estratégia estatística o perceba como relevante. Outros documentos da esfera inquisitória, tais como o relatório de inteligência policial, ou o próprio inquérito policial, relegam maior relevância às pessoas e suas identificações. O laudo pericial focaliza o corpo de delito, o *modus operandi*, e a materialização do crime. Assim, a abordagem estatística não consegue, na maioria dos casos, extrair o nome de indiciados e parceiros de um laudo.

Caberia, então, à análise linguística superar essa deficiência e se responsabilizar por essa extração. Como já discutido na Seção 3.3, sobre os instrumentos desta pesquisa, Nirenburg e Raskin (2004) ensinam que a análise semântica demanda 4 (quatro) suportes informacionais para ser realizada por completo. Deles, 2 (dois) são dependentes da linguagem: o conjunto de léxicos e o conjunto onomástico. Os outros 2 (dois) independentes: o repositório de fatos e a ontologia. O conjunto onomástico é precisamente uma coleção de substantivos próprios para detecção de nomes, uma vez que é claro que tal informação não cabe na ontologia. A arquitetura é proposta de forma que as entidades detectadas no texto que forem identificadas no conjunto onomástico sejam selecionadas como descritores do documento. Porém, esse suporte não é construído nesta pesquisa, o ferramental desenvolvido não implementa essa propriedade e o experimento não é realizado com esse requisito. Assim, neste contexto, o nome do indiciado só é selecionado como descritor do documento para aqueles laudos onde a abordagem estatística é capaz de selecioná-los por ocorrência.

A Figura 12 resume a arquitetura proposta para ilustração e fornecimento de uma visão geral do processo.

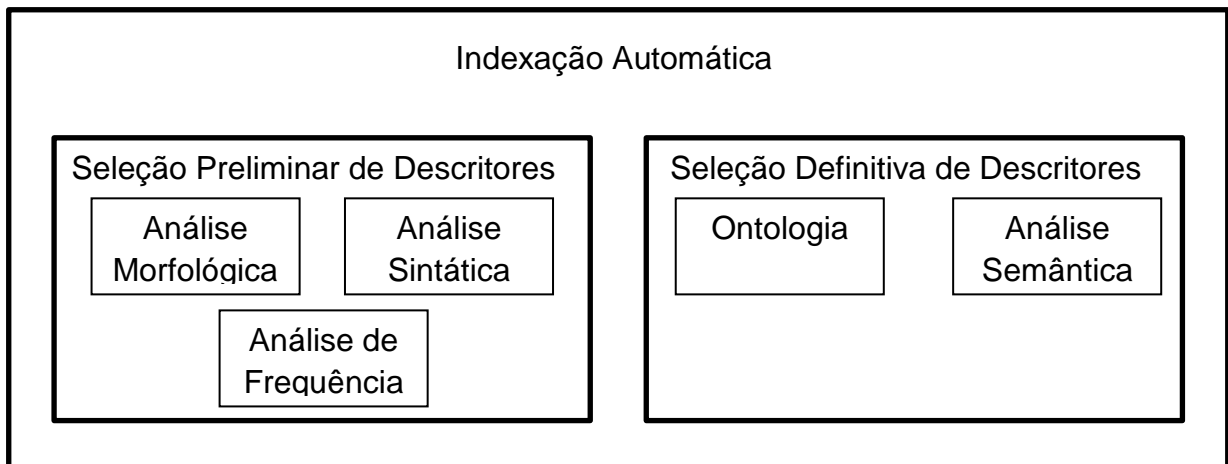


Figura 12 – Proposta de arquitetura de IA.

4.2. AVALIAÇÃO DA ARQUITETURA

Concluído o processo de IA propriamente dito, uma avaliação da qualidade da RI é realizada na base da amostra. Primeiramente há que se discutir o que são os crimes cibernéticos de que se compõem os laudos periciais da PF. Os crimes informáticos, de acordo com classificação da PF, são divididos em três categorias. A primeira qualifica os crimes puros ou próprios. Ela abrange condutas criminosas realizadas por computador e consumadas no espaço cibernético em que um recurso de informática é o alvo da ação. A invasão de uma rede de computadores, ou a produção e distribuição de vírus de computadores são exemplos desse tipo de crime. Já a segunda categoria define os crimes impuros ou impróprios. Nela os crimes são praticados por meio de computadores, mas o prejuízo ao bem jurídico tutelado ocorre fora do espaço cibernético. O furto de valores pela Internet ou a divulgação de material pornográfico infantil pela Internet bem exemplificam a segunda categoria.

A terceira categoria, por fim, trata dos crimes mediatos, indiretos ou incidentais. O computador por meio do qual uma quadrilha de traficantes de entorpecentes controla suas rotas de distribuição, contabiliza suas finanças, ou organiza seus contatos e comunicação também é corpo de delito que exige um exame pericial de informática, sendo classificado nessa terceira categoria. Percebe-se, dessa forma, que qualquer crime fora do espaço cibernético, ou seja, no mundo real, que em princípio não tenha qualquer conotação ou objetivo tecnológico, porém

que se utilize de um computador e da Internet para ser consumado, ainda assim tem relação mediata com crimes cibernéticos, e demanda perícia de informática. Nota-se, então, que a perícia de informática é evidentemente a mais demandada da PF, uma vez que a sociedade da informação e do conhecimento também abarca as organizações criminosas. Ademais, atualmente muito pouca coisa, lícita ou ilícita, é efetivamente realizada sem o apoio de um computador ou, no mínimo, de um *smartphone*.

Postas essas considerações, torna-se claro que, presentemente, qualquer crime, em potencial, demanda uma perícia de informática. Para realização deste experimento são contabilizados e catalogados os laudos da amostra que tratam das temáticas criminosas descritas no Quadro 4, assim como suas respectivas quantidades de documentos.

Temática Criminosa	Quantidade de Laudos
Ataque de negação de serviço	25
Compra de voto	9
Corrupção	236
Criação de <i>botnet</i>	16
Disseminação de programa malicioso	34
Evasão de divisas	58
Exploração de jogo de azar	89
Fraude bancária	355
Fraude em licitação	43
Fraude previdenciária	28
Injúria	6
Invasão de dispositivo informático	239
Lavagem de dinheiro	168
Pedofilia	445
Sonegação fiscal	121
Tráfico de entorpecente	219
Tráfico de influência	21
Violação de direito autoral	218
'Não classificado'	372
TOTAL	2.702

Quadro 4 – Quantidade de laudos da amostra por temática criminosa.

É de se notar, por conseguinte, algumas propriedades dos totais apresentados. Primeiramente, a soma total das quantidades de laudos por temática criminosa, 2.702 (dois mil, setecentos e dois) é maior do que a quantidade original

de documentos da amostra, qual seja 2.285 (dois mil, duzentos e oitenta e cinco) laudos de crimes cibernéticos. Isso ocorre porque um laudo pode tratar de mais de uma temática criminosa concomitantemente.

Além disso, há de se discutir que não se estão selecionando crimes propriamente ditos, no entanto temáticas criminosas, porque um crime, como definido no CP, é composto por uma atividade típica, antijurídica e imputável. A maioria dos crimes cibernéticos ainda não tem previsão legal, o que dificulta a efetiva punição de criminosos. A lei 12.737 de 30 de novembro de 2012, chamada 'Lei Carolina Dieckmann', sancionada depois do clamor popular decorrente do vazamento de fotos da atriz, acelerou o processo de tipificação do crime de invasão de dispositivo informático (BRASIL, 2013). Várias outras condutas criminosas, tais como o ataque de negação de serviço, ou a criação de *botnets*, por exemplo, ainda não têm a sua tipificação legal. Não podem, portanto, ser denominados crimes.

Outra observação remete à classe 'Não classificado'. Infelizmente há uma dualidade de opinião entre os operadores do Direito Penal que ainda não é um consenso. A primeira vertente acredita que a análise pericial é um processo estritamente científico, ou seja, prescinde de conhecimento do contexto onde a ação criminosa ocorreu e deve focalizar estritamente o corpo de delito, para identificação de materialidade e autoria. A segunda vertente pondera que o conhecimento contextual tem potencial para melhorar a qualidade do estudo científico das evidências criminosas. Os requisitantes de laudos periciais, tanto delegados de polícia, membros do MP, ou juízes de direito, que se alinham ao primeiro grupo não permitem o acesso aos dados da investigação ou processo judicial aos peritos. Os quesitos são apresentados aos PCF apenas objetivamente. Um exemplo é a solicitação de 'extração de todos os arquivos que contenham determinadas palavras chaves de algum dispositivo de armazenamento (disco rígido, *pen drive*, disquete) que seja uma evidência digital'. Já os operadores que valorizam a segunda vertente interagem diretamente com os peritos criminais para discussão dos eventos e levantamento das informações. Os quesitos podem ser construídos conjuntamente, e a tarefa de investigação e levantamento de provas é mais iterativa.

Não há consenso no Direito Brasileiro sobre qual é a melhor abordagem, mais correta, efetiva ou, por fim, mais justa. Vários argumentos jurídicos e pragmáticos, favoráveis e contrários, há para cada um dos flancos, e essa é uma discussão que ainda não apresenta indícios de proximidade de conclusão. A implicação deste fato para a perícia criminal da PF é que há uma quantidade considerável de laudos periciais em que não se sabe efetivamente qual é o crime ou temática criminosa que se está discutindo. Além disso, não é possível avaliar se os laudos construídos por extração de palavras chaves da evidência digital capturam todas as informações relevantes para a investigação. A CI, em seus estudos de RI, tende a perceber que a extração não indexada por palavra chave perde revocação, o que pode ser a diferença entre a efetiva solução de um crime ou sua não conclusão.

Percebe-se, destarte, que há muitos laudos periciais de informática que não se sabe do que efetivamente tratam. A grande quantidade de documentos desse tipo na base de dados da PF, e por extensão na amostra extraída, ainda apresenta uma incógnita de tratamento para a perícia criminal.

Em relação às quantidades de laudos por assunto, por fim, não há surpresas quanto à atuação da PF. Como polícia judiciária da União, a PF tem jurisdição e atuação preconizada pela Constituição, e se responsabiliza por ações em âmbito nacional. Assim, os maiores totais são apresentados para temáticas criminosas de foco da PF. A primeira delas são os crimes de colarinho branco, compostos por corrupção com 236 (duzentos e trinta e seis) laudos, lavagem de dinheiro com 168 (cento e sessenta e oito) documentos, e sonegação fiscal com 121 (cento e vinte e um) relatórios periciais.

Já as fraudes bancárias apresentam uma grande quantidade de laudos, com 15.53% (quinze ponto cinquenta e três por cento) do total de documentos da amostra. Isso se justifica pela parceria entre a PF e a Caixa, chamada Projeto Tentáculos. As fraudes bancárias, mormente aquelas realizadas pela Internet, são bastante mais frequentes do que publicadas na mídia, ou denunciadas para as instituições policiais, uma vez que a sensação de segurança é muito relevante para a regra de negócio das instituições financeiras. Isso dificulta o efetivo trabalho

policial. A Caixa, por se tratar de um órgão público, tem um convênio com a PF para análise proativa e reativa de segurança cibernética, o que evidentemente aumenta a quantidade de laudos e inquéritos policiais sobre o assunto.

O maior número de documentos, contudo, encontra-se nas atividades criminosas de pedofilia, corrupção de menores, e pornografia infantil, com 445 (quatrocentos e quarenta e cinco) laudos. Esse crime apresenta, infelizmente, uma parte considerável do tráfego mundial da Internet, e o esforço de instituições de segurança para desmonte de quadrilhas internacionais cresce anualmente. O tráfico de entorpecentes, por outro lado, embora apresente relação apenas mediata com crimes cibernéticos, apresenta o número de 219 (duzentos e dezenove) laudos. Em sua quase totalidade, esses documentos tratam de extração de informação de telefones e *smartphones* apreendidos em posse de traficantes.

Por fim, o último crime que se deseja discutir é o de violação de direito autoral. Esse é o crime onde normalmente é enquadrada a atividade de pirataria, que quando é realizada em âmbito interestadual ou internacional é responsabilidade da PF. Há uma quantidade razoável de laudos desse assunto, representando 9.54% (nove ponto cinquenta e quatro por cento) do total. O Brasil é um país criticado em órgãos internacionais pela alta ocorrência de pirataria, tanto física quanto cibernética.

Postas estas considerações, o Quadro 5 apresenta o cálculo do índice F para a RI de algumas consultas submetidas à base da amostra. O objetivo é exemplificar a utilização da arquitetura proposta e discutir os resultados. As linhas sombreadas dividem esses parâmetros de consultas nas temáticas criminosas descritas no Quadro 4. Assim como apresentado na Seção 2.3, as colunas da tabela se descrevem por:

RP : quantidade de documentos relevantes recuperados pela pesquisa

P : quantidade total de documentos recuperados pela pesquisa

R : quantidade total de documentos relevantes na base de dados

F : índice F

Consulta à Base da Amostra	RP	P	R	F
Ataque de negação de serviço				
negação de serviço	25	25	25	1.000000
congestionamento de servidor	25	26	25	0.980392
Compra de voto				
compra de voto	9	9	9	1.000000
eleição	9	9	9	1.000000
sufrágio	9	9	9	1.000000
Corrupção				
corrupção	236	487	236	0.652835
enriquecimento ilícito	236	308	236	0.867647
Criação de botnet				
botnet	15	15	16	0.967742
Disseminação de programa malicioso				
disseminação de programa malicioso	27	27	34	0.885246
vírus de computador	34	34	34	1.000000
Evasão de divisas				
evasão de divisas	58	59	58	0.991453
banco internacional	58	65	58	0.943089
Exploração de jogo de azar				
jogo de azar	89	89	89	1.000000
caça-níquel	73	73	73	1.000000
máquina eletrônica	73	84	73	0.929936
contabilidade de jogo	89	89	89	1.000000
programa de jogo	89	89	89	1.000000
cassino	25	25	89	0.438596
Fraude bancária				
fraude bancária	355	403	355	0.936675
cartão de crédito	84	84	84	1.000000
pagamento de conta	271	314	271	0.926496
Fraude em licitação				
fraude em licitação	43	43	43	1.000000
termo de referência	42	42	43	0.988235
projeto básico	42	42	43	0.988235
Fraude previdenciária				
fraude previdenciária	28	31	28	0.949153
inss	28	28	28	1.000000
Injúria				
injúria	6	6	6	1.000000
difamação	6	6	6	1.000000
Invasão de dispositivo informático				
invasão de dispositivo informático	239	285	239	0.912214
invasão de rede de computador	8	8	8	1.000000
invasão de computador pessoal	231	277	231	0.909449
Lavagem de dinheiro				
lavagem de dinheiro	168	197	168	0.920548
Pedofilia				
pedofilia	445	447	445	0.997758
pornografia infantil	445	445	445	1.000000

sexo	445	447	445	0.997758
Sonegação fiscal				
sonegação fiscal	121	137	121	0.937984
Tráfico de entorpecente				
tráfico de entorpecente	97	97	219	0.613924
droga de abuso	83	83	219	0.549669
Tráfico de influência				
tráfico de influência	21	26	21	0.893617
Violação de direito autoral				
violação de direito autoral	218	241	218	0.949891
pirataria	218	241	218	0.949891
cópia pirata	218	241	218	0.949891

Quadro 5 – Cálculo do índice F para consultas submetidas à base da amostra.

A escolha dos parâmetros de consulta à base da amostra foi realizada a partir das temáticas criminais selecionadas para análise, conforme apresentado no Quadro 4. Além disso, alguns quesitos dos laudos também foram eleitos, uma vez que isso remete a como uma base de dados de laudos periciais é pesquisada por usuários finais requisitantes. Percebe-se, em uma análise geral, que a revocação das pesquisas apresenta um resultado excelente, com praticamente 100% (cem por cento) de cobertura na recuperação da base. Evidentemente isso é reflexo da seleção dos parâmetros de consulta. No processo iterativo do desenvolvimento da pesquisa, e em detalhe na construção da ontologia, caso um parâmetro não retorne resultado satisfatório, basta acrescentá-lo em uma relação ontológica válida, e executar a reindexação automática da base. Nesse aspecto, a ontologia é desenvolvida e atualizada durante todo o processo, o que é corroborado pela disciplina de engenharia de ontologias, discutida na Seção 2.1, quanto à necessidade de manutenção da modelagem. Para exemplificação da arquitetura proposta, isso atende completamente o objetivo, uma vez que o benefício da análise semântica para o processo de IA se revela.

Vários casos emergem da submissão de consultas à base da amostra e demandam discussão. A taxa de 100% (cem por cento) de acerto na precisão e revocação da pesquisa por ‘negação de serviço’, por exemplo, é reflexo da estabilidade do conceito, sempre utilizado em qualquer laudo que trate do assunto. Já a pesquisa por ‘congestionamento de servidor’, que é um termo mais específico que não se encontra em nenhum dos laudos da amostra, demonstra os benefícios da arquitetura semântica. Não se pode afirmar que ‘negação de serviço’ é sinônimo

de 'congestionamento de servidor'. A relação entre eles é mais próxima de causa e efeito. O único laudo que é recuperado da base, e consequentemente polui a resposta, é um documento cujo assunto trata de pedofilia e discute a captura de pacotes em tráfego de rede. Esse laudo apresenta o termo 'congestionamento de tráfego' reiteradamente. A análise semântica considera que 'congestionamento de tráfego' é sinônimo de 'congestionamento de serviço de rede' e, por extensão taxonômica, 'congestionamento de serviço'. Assim, o objeto que atua sobre o 'serviço' é o mesmo que atua sobre o 'servidor' e a arquitetura de IA recupera o documento. Essa não é uma análise semântica exatamente correta, porém a mínima poluição do resultado representa um grande ganho frente à indexação exclusiva por palavra-chave, a qual ignoraria todos os documentos.

Já o caso da pesquisa por 'corrupção' apresenta poluição considerável no resultado da precisão. Analisando-se detidamente os documentos recuperados, verifica-se uma grande ocorrência de casos de pedofilia que contenham o termo 'corrupção de menores'. O objetivo inicial é levantar os documentos que tratem de corrupção no estrito sentido de crimes do colarinho branco. Conclui-se, então, que o parâmetro de pesquisa não é adequado, porque o termo 'corrupção' é muito geral, o qual abrange muitas interpretações semânticas e é utilizado diferentemente em muitos contextos. O termo 'enriquecimento ilícito', por outro lado, é mais específico e apresenta melhora no cálculo do índice F. Nesse caso, os documentos que poluem o resultado da pesquisa são aqueles que tratam de fraudes bancárias, o que pode, em alguns casos, ser considerado enriquecimento ilícito.

As pesquisas que tratam da temática criminosa de evasão de divisas proporcionam bom resultado para o cálculo do índice F. No caso particular do parâmetro 'banco internacional', o mesmo é selecionado propositalmente uma vez que a ocorrência nos documentos é mais tradicional para os termos 'instituição bancária internacional', 'instituição bancária americana', 'instituição bancária europeia', 'instituição financeira estrangeira' ou combinações dessas unidades lexicais. O conceito 'banco internacional' contempla todos esses parâmetros, e o resultado da pesquisa indica pequena poluição.

Na temática criminosa de exploração de jogo de azar, dos 89 (oitenta e nove) laudos detectados, 73 (setenta e três) tratam de máquinas caça-níqueis e 16 (dezesesseis) de estabelecimentos para modalidades de jogos de baralho, tais como pôquer, caixeta ou *blackjack*. A consulta pelo termo ‘cassino’ retorna resultado bastante desfavorável, tanto na revocação quanto na precisão, porque o conceito só é utilizado nos laudos que tratam de casas de jogos de cartas. Todavia o motor semântico não consegue detectar que um estabelecimento que contenha uma máquina caça-níquel também se enquadra no conceito de um cassino. Não se encontra um local adequado para acrescentar à ontologia que um bar que tenha uma máquina eletrônica de jogo é também um cassino. Assim, a pesquisa por esse termo não é adequadamente respondida. Isso corrobora a tese deste estudo no tocante a que a qualidade da análise semântica depende do conhecimento que é possível ser formalizado na ontologia.

A ontologia é o repositório de todas as relações semânticas que são estabelecidas entre os conceitos. Essas relações, no entanto, precisam estar explicitamente definidas para que o motor semântico realize suas inferências a partir delas. Se uma relação não for constituída, seja por impossibilidade do modelo ou por inépcia do processo de engenharia, a análise semântica é prejudicada. O PLN realizado nos pressupostos da arquitetura proposta por esta tese não é capaz de preencher lacunas do conhecimento formalizado no esquema de representação. Isso, como já discutido anteriormente, é foco para pesquisas na área de aprendizado automático de ontologias.

Já em relação à temática criminosa de fraude bancária, o resultado de consulta é levemente poluído por alguns laudos que tratam exclusivamente de evasão de divisas, por causa da reutilização de mesma terminologia e pela proximidade semântica dos dois assuntos. Conquanto a pesquisa por ‘cartão de crédito’ retorne resultado perfeito, a pesquisa por ‘pagamento de conta’ apresenta algum desvio na precisão quanto aos documentos que tratam de corrupção. Alguns laudos descrevem eventos de terceiros efetuando pagamento de contas particulares de agentes públicos, o que não tem amparo legal, e pode se tratar de crime de corrupção. Essa informação se mistura às fraudes bancárias onde dados de contas correntes são capturados por ataques cibernéticos para realização de pagamentos

de boletos bancários de terceiros. Essa atividade também muito se aproxima do crime de invasão de dispositivo informático, mormente quanto à invasão de computadores pessoais para extravio de informações bancárias. Há um grande número de laudos que tratam dos dois assuntos concomitantemente. Nesse caso também se percebe que a escolha do parâmetro de consulta não é precisamente amoldada para a informação que se deseja levantar.

O caso de pedofilia é emblemático nesta exemplificação de aplicação do ferramental desenvolvido para a arquitetura proposta. Primeiramente porque esse é um crime hediondo com potencial devastador para o futuro das vítimas. Segundo porque a ocorrência deste crime no Brasil é alarmante, e as instituições de segurança pública, em particular a PF, precisam discutir e atacar esse assunto da forma mais efetiva possível. As pesquisas por 'pedofilia' e 'sexo' retornam resultado quase perfeito, com apenas 2 (dois) documentos fora do assunto. Ambos os laudos tratam de injúria e difamação, e em ambos os casos fotos ou vídeos de imagens de ato sexual foram divulgados, iludindo a recuperação. Já a pesquisa por 'pornografia infantil' retorna resultado máximo para o índice F, sendo que os mais variados termos constam nos laudos, tais como 'sexo com criança', 'sexo com adolescente', 'registro de atividade sexual com criança', 'imagem pornográfica de adolescente', entre várias outras.

A pesquisa realizada no sistema Criminalística da PF, como já discutido no Capítulo 3, de metodologia, retorna um resultado deficiente, pois não há indexação da base de dados. Um exemplo de consulta textual na base demonstra esse prejuízo. Embora tenha sido identificado que haja 445 (quatrocentos e quarenta e cinco) documentos que tratem do crime de pedofilia na base da amostra, a consulta por 'pedofilia' no sistema Criminalística, filtrando-se as datas de emissão de laudos para o mesmo período, qual seja o primeiro trimestre de 2012, para todas as unidades do País, retorna apenas 203 (duzentos e três) documentos. Já a pesquisa por 'pornografia infantil', 225 (duzentos e vinte e cinco) laudos. O melhor resultado de pesquisa é obtido por meio do parâmetro 'sexo crianças adolescentes', com 363 (trezentos e sessenta e três) documentos. Mesmo que tais resultados de pesquisa apresentem precisão ideal, o que não pode ser afirmado uma vez que não é realizada detida avaliação dos documentos retornados, ainda assim o ônus para

melhoria da revocação das consultas é do usuário final, na seleção de seus parâmetros de pesquisa. O ferramental desenvolvido a partir da arquitetura de IA proposta realiza a indexação do conceito, e para esse exemplo, em particular, o resultado é consideravelmente bom.

As pesquisas relacionadas a tráfico de drogas, por sua vez, não apresentam resultado apropriado. O problema se refere ao fato de que praticamente todos esses laudos são exclusivamente de análise de telefones e *smartphones* para extração de agenda de contatos e ligações realizadas e recebidas. Há vários desses documentos em que sequer é possível a identificação da temática. Aqueles em que isso é viável normalmente o são devido a registros de mensagens de correio eletrônico ou SMS. Esses registros são comumente mascarados por siglas e metáforas criadas por traficantes, que são de difícil, senão impossível, interpretação semântica automática.

As consultas ao crime de violação de direito autoral, por fim, se beneficiam da indexação do conceito de pirataria para recuperar quaisquer laudos da base que contenham os termos 'cópia pirata', 'cópia ilegal', '*software* sem licença', 'aplicativo com licenciamento irregular', entre outros. O resultado de pesquisa é levemente poluído por documentos referentes a invasão de dispositivo informático uma vez que as ferramentas utilizadas para perpetração da atividade criminosa normalmente são *softwares* pirateados, o que obrigatoriamente consta dos respectivos laudos.

O cálculo do índice F para o resultado das consultas à base da amostra apresenta um resultado satisfatório. A média simples das pesquisas realizadas é igual a 0.929246 (zero ponto novecentos e vinte e nove, duzentos e quarenta e seis), enquanto a média ponderada pela quantidade de documentos relevantes da base é igual a 0.906066 (zero ponto novecentos e seis, zero sessenta e seis). Considerando-se que a pontuação máxima para o índice F é igual a 1 (um), há perda de menos de 10% (dez por cento) de erro experimental, o que é um resultado significativo. Evidentemente há que se considerar que esse mesmo autor constrói a ontologia e submete pesquisas à base, o que por si só já apresenta um viés. Por outro lado, a qualidade da interpretação semântica aumenta com o aprendizado do

repositório, ou seja, a arquitetura proposta preconiza que a construção da ontologia deve ser realizada a partir dos documentos que se deseja organizar para que o conhecimento formalizado reflita a informação contida na base de dados.

Concluindo, vários outros resultados de pesquisa demandam a devida análise e discussão. Não são tratados todos os casos, entretanto, uma vez que o objetivo é exclusivamente exemplificar a utilização da arquitetura proposta a partir do ferramental desenvolvido. Assim, não é possível, nem tampouco se procurou alcançar, a extensão do resultado da aplicação da implementação, até porque a amostra não probabilística extraída não apresenta quantidade suficiente de documentos para tanto. O que se percebe até então são os benefícios que a análise semântica oferece na seleção automática de descritores e, por extensão, na qualidade da RI.

Conclui-se, portanto, esta pesquisa avaliando os objetivos alvitados. O objetivo geral do trabalho se descreve por propor uma arquitetura de IA de documentos não estruturados em idioma português. A proposta de arquitetura aprofunda-se ao nível semântico de PLN utilizando uma ontologia como esquema de representação do conhecimento. As estratégias selecionadas para PLN são híbridas, com métodos estatísticos e linguísticos trabalhando concomitantemente para persecução dos resultados. Para análise léxica e extração preliminar dos descritores para IA são empregados métodos estatísticos. Para a análise sintática, semântica e extração final dos índices são aplicados métodos linguísticos.

A arquitetura proposta por este trabalho atinge este objetivo. Detalham-se os artefatos que devem ser construídos e como devem ser aproveitados para execução dos procedimentos, os quais culminam na seleção automática de descritores para documentos em língua portuguesa. Por meio de uma exemplificação utilizando ferramental implementado sobre uma base de dados de laudos periciais de crimes cibernéticos da PF, é possível levantar os benefícios advindos da proposta de arquitetura. O maior deles trata da indexação de conceitos, extraídos por análise semântica, em contraponto à indexação de palavras. Os resultados de RI para consultas à base demonstram as vantagens na abrangência e cobertura das pesquisas.

A construção do referencial teórico desta pesquisa resulta em uma revisão sobre o quadro teórico de PLN, em nível geral independente da linguagem, e em nível específico para o idioma português do Brasil. Essa é uma das contribuições inéditas desta investigação considerando a escassez de trabalhos sobre PLN lusitano, sobretudo em nível semântico. Ademais, essa é uma área de pesquisa que tem sido bastante investigada, no Brasil, pela CC e pela Linguística. A perspectiva da CI sobre o assunto é uma contribuição deste trabalho para a área.

Já em relação a seu caráter pragmático, o estudo atende o primeiro objetivo específico quanto ao desenvolvimento e integração de ferramental computacional para PLN do idioma português em nível semântico. Essa

implementação é também uma das contribuições originais da tese, a qual é posta à prova em uma base de dados, o que cumpre o terceiro objetivo. A análise dos resultados indica algumas vantagens que a arquitetura proposta oferece para a RI, principalmente quanto à qualidade dos índices de revocação e precisão de consultas.

Quanto ao segundo objetivo específico, por fim, qual seja a construção de uma ontologia de aplicação para organização e representação do conhecimento de um domínio, a ontologia desenvolvida para suporte à análise semântica é uma das contribuições inéditas da investigação. A ontologia de aplicação organizada tem estrutura de ontologia de alto nível, uma vez que seu objetivo é a modelagem semântica da linguagem. A composição, ou formato, da ontologia pode ser aplicada a qualquer domínio, enquanto a ontologia construída neste trabalho, qual seja da área de crimes cibernéticos, pode ser aplicada a quaisquer análises semânticas desse contexto.

Por outro lado, em relação às limitações desta pesquisa, há que se discutir o tamanho reduzido da amostra selecionada para experimentação do ferramental desenvolvido. Uma porção de 2.285 (dois mil, duzentos e oitenta e cinco) documentos não é suficiente para estender os resultados alcançados para qualquer base de dados. Um recorte probabilístico de maior volume é necessário para tanto. Embora a variância da amostra tenha se revelado pequena, ainda assim a estratégia de amostragem não pode ser considerada probabilística, pois apenas laudos dos 3 (três) primeiros meses do ano de 2012 foram extraídos, de forma não aleatória.

Igualmente, há de se considerar que a base de dados selecionada para aplicação do ferramental não é de domínio público. Isso não permite o confronto dos resultados ou a repetição do experimento, o que se revela uma limitação do trabalho. Para se desdobrar os resultados, é necessário avaliar o ferramental em uma base maior e, além disso, publicamente acessível. Como o objetivo é apenas exemplificar a utilização das ferramentas desenvolvidas, a base de dados atende às expectativas.

Outra limitação trata da análise semântica de PLN. Apenas 2 (dois) dos 4 (quatro) suportes informacionais para extração de significado estão disponíveis ou são construídos nesta pesquisa. Eles são o conjunto de léxicos, que é representado pela floresta sintática, e a ontologia, que é desenvolvida. Os outros 2 (dois), quais sejam o repositório de fatos e o conjunto onomástico não são produzidos, o que efetivamente prejudica a extração de significado textual. A seleção dos nomes de indicados como descritores no processo de IA dos laudos periciais bem exemplifica a falta que o conjunto onomástico representa para a análise semântica de PLN.

Percebe-se, também, que o núcleo da análise semântica da arquitetura proposta consiste na ontologia construída. O custo de construção dessa ontologia, contudo, pode tornar a arquitetura inaplicável. Uma significativa parte de todo o esforço de realização deste estudo consiste no desenvolvimento da ontologia. Isso já é esperado, consistente com a literatura e com outras pesquisas correlatas da área. Todavia impressiona o quanto o processo é penoso e, sob uma ótica de engenharia, o custo de produção tem de ser cuidadosamente avaliado para análise do benefício alcançado. Centralizar a arquitetura proposta em artefato de tamanha envergadura é uma limitação da tese.

Conquanto a exemplificação da utilização do ferramental desenvolvido tenha apresentado índices com valores altos, há de se notar que a qualidade do texto do laudo pericial pode estar maquiando o resultado alcançado. Como já discutido anteriormente, o laudo pericial é um documento rigoroso, com macroestrutura textual bem definida e morfologia e sintaxe absolutamente corretas. A realização de PLN em documento tão bem organizado é facilitada pela qualidade do mesmo. Dessa forma, uma limitação desta pesquisa se apresenta pela não realização de testes em bases de dados diferentes. Apenas se indicia, a priori, como a arquitetura proposta ou as ferramentas construídas vão se comportar em outro ambiente. Essa aplicação seria uma contribuição para a extensão dos resultados da pesquisa, no entanto a mesma não é realizada.

O pesquisador, por fim, por mais cuidado que tenha para não carregar os resultados de quaisquer vieses, é responsável por toda a cadeia de procedimentos. Isso engloba desde o desenvolvimento do ferramental, mormente a construção da

ontologia, até a escolha de parâmetros de pesquisa para submissão e avaliação de RI. Assim, é difícil garantir a não obliquidade, o que é uma limitação do estudo. Caso as consultas à base da amostra fossem realizadas por outros pesquisadores ou voluntários, isso já traria novas perspectivas e percepções à análise da arquitetura proposta.

Vários trabalhos futuros emergem das conclusões desta pesquisa. O primeiro deles é, evidentemente, a efetiva implementação da arquitetura de SI proposta, e a implantação em ambiente de produção do ferramental computacional construído. Os resultados alcançados por esta pesquisa permitem apenas entrever as vantagens da utilização da arquitetura. A concretização dos resultados, entretanto, só pode ser identificada em ambiente real de utilização por usuários finais.

Outro trabalho futuro trata da evolução da arquitetura proposta. O NLTK possui um módulo para tratamento de GLC probabilística, denominado 'parse_pcfg'. Essa abordagem parece ser interessante para melhoria das intuições de ambiguidades na análise sintática. O *parse* probabilístico é uma abordagem híbrida por si só. A um método linguístico de produção da árvore de derivação é somado uma estratégia estatística de probabilidade de execução de uma regra de *parse*. A agregação desses pressupostos à arquitetura proposta, assim como a implementação das ferramentas e efetiva aplicação é um trabalho que pode ser desenvolvido para medição e comparação com os resultados atuais.

Ainda na evolução da arquitetura, vários trabalhos futuros surgem na utilização de outros algoritmos de derivação para análise sintática. O objetivo é possibilitar a comparação de taxas de acerto na classificação linguística. O algoritmo de Earley, por exemplo, possui uma extensão que permite a associação de pressupostos semânticos às regras sintáticas. Como discutido na Seção 2.2.4, isso permite a criação de uma relação regra a regra entre a sintaxe e a semântica da linguagem, o que pode oferecer bom resultado nas análises. A modificação da arquitetura para essa abordagem, ou quaisquer das outras elencadas no levantamento, é um exercício que potencialmente valida a arquitetura proposta ou propõe melhorias a ela.

Já em relação ao experimento propriamente dito, um trabalho futuro se descreve na realização daquele em uma base de dados maior. A amostra selecionada é muito restrita, e não permite a extensão dos resultados alcançados. Um conjunto de laudos periciais de maior volume, inclusive de outras áreas técnicas, garante a validação do ferramental desenvolvido nesse ambiente. Para a PF, em particular, e para qualquer órgão da administração pública federal, de maneira geral, esse é um estudo de caso de grande agregação de valor para a instituição, para melhor estruturação de seus SI.

Outrossim, a realização do experimento em uma base de dados diferente possibilita a avaliação da efetiva investigação semântica da linguagem. A escolha de outro conjunto de documentos, com macroestrutura e estilos de elocução dessemelhantes, é uma prova para utilização e evidência dos benefícios da arquitetura proposta. A extensão do quanto a arquitetura de SI tem de passar por outro processo de treinamento para alcançar níveis análogos de qualidade na RI vai demonstrar o quanto a proposta é aplicável a diferentes cenários.

O último trabalho futuro que se deseja propor, por fim, trata da utilização do motor semântico de PLN para outras finalidades. Já se discutiu o custo elevado que a construção da ontologia apresenta. Assim, há de se avaliar o quanto o aprofundamento até a análise semântica é, de fato, válido ou necessário para IA. Não se discute a importância da área de pesquisa em IA, a qual ainda tem vários problemas abertos e não apresenta resultados conclusivamente estabelecidos. Ocorre, porém, que é possível que análises linguísticas aprofundadas somente até o nível sintático, associadas a esquemas de representação do conhecimento menos elaborados, tais como um tesouro, possam alcançar resultados não tão expressivos, mas ainda assim satisfatórios, a um custo grosseiramente menor.

Sob este aspecto, há de se levantar novos fins para o motor de PLN em nível semântico. A TA é uma aplicação de PLN que demanda intrinsecamente tal aprofundamento linguístico. Essa é uma área de pesquisa com vários problemas abertos que tem uma demanda crescente considerando a produção global de conhecimento, em uma infinidade de idiomas diferentes, e o repositório universal da

Internet. Além dessa, os sistemas de pergunta e resposta apresentam potencial para modificar o paradigma de busca e acesso de informação. Ao contrário de buscar e recuperar informações sobre determinado assunto, que é o *modus operandi* atual, o objetivo é responder objetivamente perguntas formuladas por usuários. Para isso, a extração de significado é fundamental. Várias outras aplicações, além desses exemplos, se beneficiam da análise semântica de PLN. O motor proposto na arquitetura desta tese pode ser utilizado, em trabalhos futuros, para suportar esses sistemas.

Concluindo, finalmente, afirma-se que PLN é um dragão. O dragão é uma clássica alegoria que a CC faz para os compiladores. Ele é a capa da mais tradicional obra da área, utilizada na formação de cientistas de computação em todo o mundo. O dragão representa a enorme dificuldade em traduzir a linguagem formal de uma linguagem de programação para a codificação binária e executável compreendida pelo processador de uma máquina. O dragão de PLN é muito maior, porquanto a linguagem natural não é livre de contexto e as gramáticas para esse tipo de linguagem não conseguem se precaver de ambiguidades. PLN no idioma português, então, representa uma tarefa de ainda maior dificuldade por causa da abissal complexidade da língua e, por extensão, de sua riqueza. Propõe-se, destarte, aos distintos pesquisados de PLN no idioma de Camões a mesma ousadia e coragem que os primeiros desbravadores lusitanos apresentaram ao singrar mares desconhecidos e enfrentar desafios para alcance de seus objetivos.

Os Lusíadas

Luís de Camões, 1572

As armas e os barões assinalados
Que, da ocidental praia lusitana,
Por mares nunca dantes navegados,
Passaram ainda além da Taprobana,
Em perigos e guerras esforçados
Mais do que prometia a força humana,
E entre gente remota edificaram
Novo reino, que tanto sublimaram;
(CAMÕES, 1980)

ALENCAR, L. F. Utilização de informações lexicais extraídas automaticamente de corpora na análise sintática computacional do português. **Revista de estudos da linguagem**, v. 19, n.1, p. 7–85, 2011.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da informação**, v. 32, n. 3, p. 7–20, 2003.

ANDERSON, J. D.; PÉREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: machine indexing, and the allocation of human versus machine effort. **Information processing and management**, v. 37, p. 255–277, 2001.

ANGELE, J.; KIFER, M.; LAUSEN, G. Ontologies in f-logic. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

ANTONIOU, G.; VAN HARMELEN, F. Web ontology language: OWL. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

ARAÚJO, A. **As pontes de Königsberg**. Disponível em: <<http://www.mat.uc.pt/~alma/escolas/pontes/>>. Acesso em: 16 jul. 2012.

BAADER, F.; HORROCKS, I.; SATTLER, U. Description logics. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval: the concepts and technology behind search**. 2. ed. Boston: Addison Wesley, 2011.

BALDWIN, T.; KIM, S. N. Multiword expressions. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

BELGHIAT, A.; BOURAHLA, M. An approach based AToM3 for the generation of OWL ontologies from UML diagrams. **International journal of computer applications**, v. 41, n. 3, p. 41–48, 2012.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with python**. Sebastopol: O'Reilly, 2009.

BLACKBURN, P.; BOS, J. **Representation and inference for natural language: a first course in computational semantics**. Stanford: CSLI Publications, 2005.

BLANCO, R.; LIOMA, C. Graph-based term weighting for information retrieval. **Information retrieval**, v. 15, p. 54–92, 2012.

BOOTH, T. L.; THOMPSON, R. A. Applying probability measures to abstract languages. **Transactions on computers**, v. C-22 , n. 5, p. 442–450, 1973.

BORGO, S.; MASOLO, C. Foundational choices in DOLCE. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

BORKO, H. Information science: what is it? **American documentation**, v. 19, n. 1, 1968.

BORKO, H. Toward a theory of indexing. **Information processing and management**, v. 13, p. 355–365, 1977.

BRASIL. **Código de processo penal e constituição federal**. 49. ed. São Paulo: Saraiva, 2009.

BRASIL. **Lei nº 12.737**, de 30 de novembro de 2012. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12737.htm>. Acesso em: 15 jan. 2013.

BROOKES, B. C. The foundations of information science. Part I. Philosophical aspect. **Journal of information science**, n. 2, p. 125–133, 1980.

BRUIJN, J.; KERRIGAN, M.; ZAREMBA, M.; FENSEL, D. Semantic web services. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

CAMARA JUNIOR, A. T. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. Dissertação de Mestrado, Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2007.

CAMÕES, L. **Os Lusíadas**. 4. ed. São Paulo: Cultrix, 1980.

CASTELLS, M. **A Sociedade em Rede**. 8 ed. São Paulo: Paz e Terra, 2005.

CHARTON, E.; TORRES–MORENO, J–M. Modélisation automatique de connecteurs logiques par analyse statistique du contexte. **Canadian journal of information and library science**, v. 35, n. 3, p. 287–306, 2011.

CHAUDIRON, S. Technologies linguistiques et modes de représentation de l'information textuelle. **Documentaliste – sciences de l'information**, v. 44, n. 1, p. 30–39, 2007.

CHELBA, C. Statistical language modeling. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.

CHERPAS, C. Natural language processing, pragmatics, and verbal behavior. **The analysis of verbal behavior**, v. 10, p. 135–147, 1992.

CHOI, F. Y. Y.; WIEMER–HASTINGS, P.; MOORE, J. Latent semantic analysis for text segmentation. **Proceedings of the conference on empirical methods in natural language processing**, Pittsburgh, p. 109–117, 2001.

CHOMSKY, N. Three models for the description of language. **Transactions on information theory**, v. 2, n. 3, p. 113–124, 1956.

CHOMSKY, N. **Aspects of the theory of syntax**. Cambridge: MIT Press, 1969.

CHUNG, E.; MIKSA, S.; HASTINGS, S. K. A framework of automatic subject term assignment for text categorization: an indexing conception-based approach. **Journal of the american society for information science and technology**, v. 61, n. 4, p. 688–699, 2010.

CIMIANO, P.; VÖLKER, J.; BUITELAAR, P. Ontology construction. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

CLARK, S. Statistical parsing. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.

CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.

CLARK, A.; LAPPIN, S. Unsupervised learning and grammar induction. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.

CONSTANT, M.; SIGOGNE, A.; WATRIN, P. La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa: évaluation de deux stratégies discriminantes. **Travaux de la 19^{ème} conférence sur le traitement automatique des langues naturelles**, Grenoble, p. 57–70, 2012.

CRESWELL, J. W. **Research design: qualitative, quantitative, and mixed methods approach**. 3. ed. Los Angeles: Sage, 2009.

CRUSE, A. **Meaning in language: an introduction to semantics and pragmatics**. 3. ed. New York: Oxford University Press, 2011.

DAELEMANS, W.; BOSCH, A. V. D. Memory-based learning. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley-Blackwell, 2010.

DALE, R. Classical approaches to natural language processing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

DI FELIPPO, A.; DIAS-DA-SILVA, B. C. O processamento automático de línguas naturais enquanto engenharia do conhecimento linguístico. **Calidoscópico**, v. 7, n. 3, p. 183-191, 2009.

ERJAVEC, T.; DZEROSKI, S. Machine learning of morphosyntactic structure: lemmatizing unknown slovene words. **Applied artificial intelligence**, v. 18, p. 17-41, 2004.

FOX, C. Computational semantics. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley-Blackwell, 2010.

FREITAS, C.; ROCHA, P.; BICK, E. Um mundo novo na floresta sintá(c)tica – o treebank do português. **Calidoscópico**, v. 6, n. 3, p. 142-148, 2008.

FURNAS, G. W.; DEERWESTER, S.; DUMAIS, S. T.; LANDAUER, T. K.; HARSHMAN, R. A.; STREETER, L. A.; LOCHBAUM, K. E. Information retrieval using a singular value decomposition model of latent semantic structure. **Proceedings of the 11th annual international conference on research and development in information retrieval**, Grenoble, p. 465-480, 1988.

GARCIA, M.; GAMALHO, P. Análise morfossintáctica para português europeu e galego: problemas, soluções e avaliação. **Linguamática**, v. 2, n. 2, p. 59-67, 2010.

- GODDARD, C.; SCHALLEY, A. C. Semantic analysis. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.
- GOLDSMITH, J. A. Segmentation and morphology. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.
- GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge acquisition**, v. 5, p. 199–220, 1993.
- GUARINO, N. Understanding, building and using ontologies. **International journal of human–computer studies**, v. 46, p. 293–310, 1997.
- GUARINO, N. Formal ontology and information systems. **Proceedings of the formal ontology in information systems**, Trento, p. 3–15, 1998.
- GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.
- GUARINO, N.; WELTY, C. A. An overview of OntoClean. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.
- GÜNGÖR, T. Part-of-speech tagging. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.
- HAGE, J.; VERHEIJ, B. The law as a dynamic interconnected system of states of affairs: a legal top ontology. **International journal of human–computer studies**, v. 51, n. 6, p. 1043–1077, 1999.
- HAJICOVÁ, E.; ABEILLÉ, A.; HAJIC, J.; MÍROVSKÝ, J.; URESOVÁ, Z. Treebank Annotation. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

- HEARST, M. A. Multi-paragraph segmentation of expository text. **Proceedings of the 32th annual meeting of the association for computational linguistics**, Las Cruces, p. 9–16, 1994.
- HEARST, M. A. Texttiling: segmenting text into multi-paragraph subtopic passages. **Computational linguistics**, v. 23, n. 1, p. 33–64, 1997.
- HEINECKE, J.; SMITS, G.; CHARDENON, C.; DE NEEF, E. G.; MAILLEBUAU, E.; BOUALEM, M. TiLT plate-forme pour le traitement automatique des langues naturelles. **Traitement Automatique des Langues**, v. 49, n. 2, p. 17–41, 2008.
- HENDERSON, J. B. Artificial neural networks. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.
- HERTEL, A.; BROEKSTRA, J.; STUCKENSCHMIDT, H. RDF storage and retrieval systems. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.
- HIPPISLEY, A. Lexical analysis. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.
- HIRST, G. Ontology and the lexicon. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.
- HITZLER, P.; PARSIA, B. Ontologies and rules. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.
- HOBBS, J. Coherence and coreference. **Cognitive science**, v. 3, p. 67–90, 1979.
- HOLLAND, G. A. Information science: an interdisciplinary effort? **Journal of documentation**, v. 64, n. 1, p. 7–23, 2008.

HOPCROFT, J. E.; MOTWANI, R.; ULLMAN, J. D. **Introduction to automata theory, languages, and computation**. 3. ed. Boston: Addison Wesley, 2006.

JACKSON, P.; AL-KOFAHI, K.; TYRRELL, A.; VACHHER, A. Information extraction from case law and retrieval of prior cases. **Artificial intelligence**, v. 150, p. 239–290, 2003.

JANSEN, B. J.; RIEH, S. Y. The seventeen theoretical constructs of information searching and information retrieval. **Journal of the american society for information science and technology**, v. 61, n. 8, p. 1517–1534, 2010.

JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of IR techniques. **Transactions on information systems**, v. 20, n. 4, p. 422–446, 2002.

JOSHI, S. D.; DESHPANDE, D. Textual requirement analysis for UML diagram extraction by using NLP. **International journal of computer applications**, v. 50, n. 8, p. 42–46, 2012.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition**. 2. ed. Upper Saddle River: Pearson Prentice Hall, 2008.

KASAMA, D. Y.; ZAVAGLIA, C.; ALMEIDA, G. M. B. Do termo à estruturação semântica: representação ontológica do domínio da nanociência e nanotecnologia utilizando a estrutura qualia. **Linguamática**, v. 2, n. 3, p. 43–58, 2010.

KURAMOTO, H. **Proposition d'un système de recherche d'information assistée par ordinateur: avec application à la langue portugaise**. Tese de Doutorado, L'Université Lumière, Lyon, 1999.

LACY, L. W. **OWL: representing information using the web ontology language**. Victoria: Trafford, 2005.

LADEIRA, A. P. **Processamento de linguagem natural**: caracterização da produção científica dos pesquisadores brasileiros. Tese de Doutorado, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

LAHTINEN, T. **Automatic indexing**: an approach using an index term corpus and combining linguistic and statistical methods. Tese de Doutorado, Department of General Linguistics, University of Helsinki, Helsinki, 2000.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

LE COADIC, Y. F. **A ciência da informação**. 2. ed. Brasília: Briquet de Lemos, 2004.

LEE, D. H.; SCHLEYER, T. Social tagging is no substitute for controlled indexing: a comparison of medical subject headings and CiteULike tags assigned to 231,388 papers. **Journal of the american society for information science and technology**, v. 63, n. 9, p. 1747–1757, 2012.

LI, D.; KWONG, C.–P. Understanding latent semantic indexing: a topological structure analysis using q-analysis. **Journal of the american society for information science and technology**, v. 61, n. 3, p. 592–608, 2010.

LIU, V.; CURRAN, J. R. Web text corpus for natural language processing. **Proceedings of the 11th conference of the european chapter of the association for computational linguistics**, Trento, p. 233–240, 2006.

LJUNGLÖF, P.; WIRÉN, M. Syntactic parsing. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

- MAIA, L. C. G. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. Tese de Doutorado, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.
- MALOUF, R. Maximum entropy models. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.
- MARCONI, M. A.; LAKATOS, E. M. **Metodologia científica**. 4. ed. São Paulo: Atlas, 2004.
- MARTINS, R. T.; HASEGAWA, R.; NUNES, M. G. V. **Curupira**: um parser funcional para a língua portuguesa. Disponível em: <<http://www.nilc.icmc.usp.br/nilc/download/nilc-tr-02-26.zip>>. Acesso em: 27 set. 2012.
- MCDONALD, R.; PEREIRA, F.; RIBAROV, K.; HAJIC, J. Non-projective dependency parsing using spanning tree algorithms. **Proceedings of the human language technology conference and conference on empirical methods in natural language processing**, Vancouver, p. 523–530, 2005.
- MELLISH, C.; PAN, J. Z. Natural language directed inference from ontologies. **Artificial intelligence**, v. 172, p. 1285–1315, 2008.
- MIHALCEA, R.; RADEV, D. **Graph based natural language processing and information retrieval**. New York: Cambridge University Press, 2011.
- MITKOV, R. Discourse processing. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.
- MOENS, M.–F. **Automatic indexing and abstracting of document texts**. Norwell: Kluwer Academic Publishers, 2000.

- MOENS, M.-F.; UYTENDAELE, C.; DUMORTIER, J. Information extraction from legal texts: the potential of discourse analysis. **International journal of human-computer studies**, v. 51, n. 6, p. 1155–1171, 1999.
- MOREIRA, W. **A construção de informações documentárias**: aportes da linguística documentária, da terminologia e das ontologias. Tese de Doutorado, Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2010.
- NEDERHOF, M.-J.; SATTÀ, G. Theory of parsing. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.
- NÉVÉOL, A.; ROGOZAN, A.; DARMONI, S. Automatic indexing of online health resources for a french quality controlled gateway. **Information processing and management**, v. 42, p. 695–709, 2006.
- NIRENBURG, S.; RASKIN, V. **Ontological Semantics**. Cambridge: MIT Press, 2004.
- NIVRE, J. Statistical parsing. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.
- NIVRE, J.; MCDONALD, R. Integrating graph-based and transition-based dependency parsers. **Proceedings of the association for computational linguistics**, Columbus, p. 950–958, 2008.
- OBERLE, D.; GRIMM, S.; STAAB, S. An ontology for software. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.
- OLIVEIRA, H. G.; SANTOS, D.; GOMES, P. Extração de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. **Linguamática**, v. 2, n. 1, p. 77–93, 2010.

PALMER, D. D. Text preprocessing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

PALMER, D. D.; HEARST, M. A. Adaptive multilingual sentence boundary disambiguation. **Computational linguistics**, v. 23, n. 2, p. 241–267, 1997.

PAN, J. Z. Resource description framework. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

PINTO, H. S.; TEMPICH, C.; STAAB, S. Ontology engineering and evolution in a distributed world using DILIGENT. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130–137, 1980.

PRATT–HARTMANN, I. Computational complexity in natural language. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.

PULGARÍN, A.; GIL–LEIVA, I. Bibliometric analysis of the automatic indexing literature: 1956–2000. **Information processing and management**, v. 40, p. 365–377, 2004.

PUSTEJOVSKY, J. The generative lexicon. **Computational linguistics**, v. 17, n. 4, p. 409–441, 1991.

RESNIK, P.; LIN, J. Evaluation of NLP systems. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.

RIBEIRO, R.; OLIVEIRA, L.; TRANCOSO, I. Using morphosyntactic information in TTS systems comparing strategies for european portuguese. **Proceedings of the 6th**

international workshop on computational processing of the portuguese language, Faro, p. 143–150, 2003.

ROBREDO, J. **Documentação de hoje e de amanhã**: uma abordagem revisitada e contemporânea da ciência da informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. Brasília: Reprint, 2005.

ROSENFELD, L.; MORVILLE, P. M. **Information architecture for the world wide web**. Sebastopol: O'Reilly, 2002.

SABAH, G. Natural language understanding, where are we going? Where could we go? **The computer journal**, v. 54, n. 9, p. 1505–1513, 2011.

SANTOS, E. S. **Uma proposta de integração de sistemas computacionais utilizando ontologias**. Dissertação de Mestrado, Departamento de Ciência da Computação, Universidade de Brasília, Brasília, 2006.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da informação**, v. 24, n.1, 1995.

SAVOY, J.; GAUSSIÉ, E. Information retrieval. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

SCHALLEY, A. C. Representing verbal semantics with diagrams. An adaptation of the UML for lexical semantics. **Proceedings of the 20th international conference on computational linguistics**, Genève, v. 2, p. 785–791, 2004.

SCHMID, H. Decision trees. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.

SHINDE, S. K.; BHOJANE, V.; MAHAJAN, P. NLP based object oriented analysis and design from requirement specification. **International journal of computer applications**, v. 47, n. 21, p. 30–34, 2012

SINHA, R.; MIHALCEA, R. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. **Proceedings of the IEEE international conference on semantic computing**, Irvine, 2007.

SOMMERVILLE, I. **Software engineering**. 9. ed. Boston: Addison Wesley, 2011.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Revista eletrônica de biblioteconomia e ciência da informação**. Edição especial 1º semestre de 2006, p. 42–59, 2006.

SOWA, J. F. Semantics of conceptual graphs. **Proceedings of the 17th annual meeting on association for computational linguistics**, Stroudsburg, p. 39–44, 1979.

STEVENS, R.; LORD, P. Application of ontologies in bioinformatics. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

SURE, Y.; STAAB, S.; STUDER, R. Ontology engineering methodology. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

SVENONIUS, E. **The intellectual foundation of information organization**. Cambridge: MIT Press, 2000.

VRANDECIC, D. Ontology evaluation. In: STAAB, S.; STUDER, R. (Ed.) **Handbook on ontologies**. 2. ed. Berlin: Springer, 2009.

WIDDOWS, D.; DOROW, B. A graph model for unsupervised lexical acquisition. **Proceedings of the 19th international conference on computational linguistics**, Taipei, 2002.

- WILLIAMS, R. V. Hans Peter Luhn and Herbert M. Ohlman: their roles in the origins of keyword-in-context/permutation automatic indexing. **Journal of the american society for information science and technology**, v. 61, n. 4, p. 835–849, 2010.
- WINTNER, S. Formal language theory. In: CLARK, A.; FOX, C.; LAPPIN, S. (Ed.) **The handbook of computational linguistics and natural language processing**. Chichester: Wiley–Blackwell, 2010.
- WU, D.; ZHANG, Y.; LIU, T. Unsupervised query segmentation using monolingual word alignment method. **Computer and information science**, v. 5, n. 1, p. 13–19, 2012.
- WURMAN, R. S. **Ansiedade da informação 2**. São Paulo: Editora de Cultura, 2005.
- XIAO, R. Corpus Creation. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.
- YAROWSKY, D. Word sense disambiguation. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.
- YUAN, X.; BELKIN, N. J. Investigating information retrieval support techniques for different information-seeking strategies. **Journal of the american society for information science and technology**, v. 61, n. 8, p. 1543–1563, 2010.
- ZHANG, T. Fundamental statistical techniques. In: INDURKHAYA, N.; DAMERAU, F. J. (Ed.) **Handbook of natural language processing**. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.
- ZOBEL, J.; MOFFAT, A. Inverted files for text search engines. **Computing surveys**, v. 38, n. 2, p. 1–56, 2006.

APÊNDICE A – SCRIPT PYTHON

```
#####  
# UnB - Universidade de Brasilia #  
# FCI - Faculdade de Ciencia da Informacao #  
# PPGCInf - Programa de Pos-Graduacao em Ciencia da Informacao #  
# Tese de Doutorado #  
# Processamento de Linguagem Natural para Indexacao Automatica SemanticoOntologica#  
# Auto Tavares da Camara Junior #  
# Script Python - 11/04/2013 #  
#####  
  
#####  
# Importacao de bibliotecas #  
#####  
  
from xml.dom import minidom  
import nltk  
from nltk import data  
from nltk import parse_cfg  
from nltk import pos_tag  
from nltk import RecursiveDescentParser  
from nltk import FreqDist  
from nltk.corpus import PlaintextCorpusReader  
from nltk.corpus import floresta  
  
#####  
# Definicao de constantes #  
#####  
  
diretorio = '/home/junior'  
diretorioLaudos = diretorio + '/laudos'  
  
#####  
# Inicializacao de Variaveis #  
#####  
  
listaDeArvoresDeParse = []  
listaPreliminarDeDescritores = []  
listaDefinitivaDeDescritores = []  
  
#####  
# Carregamento da gramatica #  
#####  
  
gramaticaCurupira = parse_cfg("""  
FRASE -> PERIODO | delimitador AADVO SUJ AADVO delimitador  
  
PERIODO -> PERIODO_COORDENADO | PERIODO_INDEPENDENTE  
PERIODO_COORDENADO -> coordenador PERIODO_INDEPENDENTE coordenador  
PERIODO_COORDENADO | coordenador PERIODO_INDEPENDENTE coordenador  
PERIODO_INDEPENDENTE  
PERIODO_INDEPENDENTE -> AADVO SUJ AADVO PREDICADO AADVO | AADVO PREDICADO AADVO  
SUJ AADVO | AADVO PREDICADO AADVO  
  
SUJ -> OSSS | SUJ_COMPOSTO | SUJ_SIMPLES  
SUJ_COMPOSTO -> coordenador SUJ_SIMPLES coordenador SUJ_COMPOSTO | coordenador  
SUJ_SIMPLES coordenador SUJ_SIMPLES  
SUJ_SIMPLES -> 'eu' | 'tu' | 'ele' | 'ela' | 'nos' | 'vos' | 'eles' | 'elas' | SN  
  
PREDICADO -> PREDVN | PREDV | PREDN  
PREDN -> SVL PSUJ  
PREDV -> verbo ODA SVTDI OI | verbo OIA SVTDI OD | verbo ODA SVTD | verbo OIA SVTI  
| ODA SVTDI OI | OIA SVTDI OD | ODA SVTD | OIA SVTI | SVTDI OD OI | SVTDI OI AP |  
SVTDI OI OD | SVTD AP | SVTD OD | SVTD OI | SVI
```


PREDVN -> ODA SVTD POBJ | ODA SVTD PSUJ | SVTI OI POBJ | SVTD OD POBJ | SVTD POBJ
 OD | SVTD OD PSUJ | SVI PSUJ

POBJ -> SADJ | SN

PSUJ -> OSSPSUJ | PSUJ_COMPOSTO | PSUJ_SIMPLES | coordenador PSUJ_SIMPLES
 coordenador PSUJ_COMPOSTO | coordenador PSUJ_SIMPLES coordenador PSUJ_SIMPLES |
 SADJ | SN

CN -> OSSCN | SP

OD -> hifen 'me' | hifen 'te' | hifen 'se' | hifen 'o' | hifen 'a' | hifen 'nos' |
 hifen 'vos' | hifen 'os' | hifen 'as' | OSSOD | OD_COMPOSTO | OD_SIMPLES
 OD_COMPOSTO -> coordenador OD_SIMPLES coordenador OD_COMPOSTO | coordenador
 OD_SIMPLES coordenador OD_SIMPLES
 OD_SIMPLES -> SN
 ODA -> 'me' | 'te' | 'se' | 'o' | 'a' | 'nos' | 'vos' | 'os' | 'as'

OI -> hifen 'me' | hifen 'te' | hifen 'se' | hifen 'lhe' | hifen 'nos' | hifen
 'vos' | hifen 'lhes' | poi 'mim' | poi 'ti' | poi 'si' | poi 'ele' | poi 'ela' |
 poi 'nos' | poi 'vos' | poi 'eles' | poi 'elas' | OSSOI | OI_COMPOSTO | OI_SIMPLES
 OI_COMPOSTO -> coordenador OI_SIMPLES coordenador OI_COMPOSTO | coordenador
 OI_SIMPLES coordenador OI_SIMPLES
 OI_SIMPLES -> poi SN
 OIA -> 'me' | 'te' | 'se' | 'lhe' | 'nos' | 'vos' | 'lhes'

AP -> OSSAP | pap SN

AADND -> OSADJ | AADND_COMPOSTO | AADND_SIMPLES
 AADND_COMPOSTO -> coordenador AADND_SIMPLES coordenador AADND_COMPOSTO |
 coordenador AADND_SIMPLES coordenador AADND_SIMPLES
 AADND_SIMPLES -> SADJ | SP
 AADNE -> SDET nucleo | SDET

AADVO -> OSADV | 'comigo' | 'contigo' | 'consigo' | 'conosco' | 'convosco' | paadv
 'mim' | paadv 'ti' | paadv 'si' | paadv 'ele' | paadv 'ela' | paadv 'nos' | paadv
 'vos' | paadv 'eles' | paadv 'elas' | AADVO_COMPOSTO | AADVO_SIMPLES
 AADVO_COMPOSTO -> coordenador AADVO_SIMPLES coordenador AADVO_COMPOSTO |
 coordenador AADVO_SIMPLES coordenador AADVO_SIMPLES
 AADVO_SIMPLES -> SADV | SP

AADVL -> 'comigo' | 'contigo' | 'consigo' | 'conosco' | 'convosco' | paadv 'mim' |
 paadv 'ti' | paadv 'si' | paadv 'ele' | paadv 'ela' | paadv 'nos' | paadv 'vos' |
 paadv 'eles' | paadv 'elas' | AADVL_COMPOSTO | AADVL_SIMPLES
 AADVL_COMPOSTO -> coordenador AADVL_SIMPLES coordenador AADVL_COMPOSTO |
 coordenador AADVL_SIMPLES coordenador AADVL_SIMPLES
 AADVL_SIMPLES -> SADV

APOSTO -> OSSAPO | SN

SADJ -> AADVL nucleo CN SADJ

SADV -> AADVL nucleo CN AADVL

SDET -> nucleo | 'cerca de' | 'perto de' | 'mais de'

SN -> pron | AADVL AADNE nucleo CN AADND | AADVL AADNE nucleo AADND

SP -> p SN | p SADJ | p SADV

SVL -> AADVL verbo verbo vl
 SVTD -> AADVL verbo verbo vtd
 SVTDI -> AADVL verbo verbo vtdi
 SVTI -> AADVL verbo verbo vti
 SVI -> AADVL verbo verbo vi

OSADJ -> ORG | ORP | SREL

```

OSADV -> ORG | ORP | subordinante PERIODO

OSS -> ORI | integrante PERIODO
OSSAPO -> PERIODO
OSSCN -> pcn OSS
OSSOD -> OSS
OSSOI -> poi OSS
OSSPSUJ -> OSS
OSSS -> OSS
OSSAP -> OSS

ORI -> PERIODO
ORP -> PERIODO
ORG -> PERIODO

SREL -> PERIODO

coordenador -> virgula conjuncao_coordenativa | virgula | ponto_e_virgula

delimitador -> reticencias | ponto_de_interrogacao | ponto_de_exclamacao |
dois_pontos | ponto_final | marcador_de_fim_de_paragrafo | marcador_de_tabulacao |
marcador_de_fim_de_linha

integrante -> conjuncao_integrante

nucleo -> adjetivo | numeral_cardinal | numeral_ordinal |
pronome_demonstrativo_variavel | pronome_indefinido_variavel | pronome_possessivo |
algarismo_arabico | adverbio | substantivo | nome_proprio |
toda_e_qualquer_palavra_desconhecida | sigla | abreviatura | numeral_multiplicativo |
numeral_fracionario | numeral_coletivo

p -> 'a' | 'ante' | 'apos' | 'ate' | 'com' | 'contra' | 'de' | 'desde' | 'em' |
'entre' | 'para' | 'per' | 'perante' | 'por' | 'sem' | 'sob' | 'sobre' | 'tras'

paadv -> p

pap -> 'de' | 'por'

pcn -> 'de' | 'em' | 'com'

poi -> 'de' | 'em' | 'com' | 'para'

pron -> pronome_demonstrativo_invariavel | pronome_indefinido_invariavel |
pronome_interrogativo | pronome_relativo

subordinante -> conjuncao_subordinativa

vi -> verbo_intransitivo

vl -> verbo_de_ligacao

vtd -> verbo_transitivo_direto

vtidi -> verbo_transitivo_direto_e_indireto

vti -> verbo_transitivo_indireto

verbo -> verbo_auxiliar | 'comecar a' | 'terminar de' | 'continuar a' | 'deixar
de'
"""

#####
# Carregamento da base de treinamento para a lingua portuguesa #
#####

data.load('tokenizers/punkt/portuguese.pickle')

#####

```

```

# Carregamento da ontologia
#####

xmldoc = minidom.parse('ontologia.owl')
listaDeConceitosOntologia = xmldoc.getElementsByTagName('Conceito')
print 'Quantidade de conceitos na ontologia: ' +
str(len(listaDeConceitosOntologia))

#####
# Leitura dos laudos
#####

listaDeLaudos = PlaintextCorpusReader(diretorioLaudos, '.*')

for laudo in listaDeLaudos.fileids():
    arquivoLaudo = listaDeLaudos.open(laudo, 'rU')

    print 'Laudo em analise: ' + laudo

    contador = 0
    for linha in arquivoLaudo:
        print 'Linha ' + str(contador) + ' do laudo: ' + linha.strip()
        contador = contador + 1

#####
# Analise morfosintatica
#####

sentencas = listaDeLaudos.sents(fileids=laudo)
for sentenca in sentencas:
    print 'Etiquetagem POS da sentenca: ' + str(pos_tag(sentenca))

    parser = RecursiveDescentParser(gramaticaCurupira)
    arvores = parser.nbest_parse(sentenca)
    for arvore in arvores:
        print 'Arvore de parse sintatica da sentenca: ' + str(arvore)
        listaDeArvoresDeParse.append(arvore)

#####
# Extracao preliminar de descritores: analise de frequencia
#####

distribuicaoFrequencia = FreqDist(listaDeArvoresDeParse)

contador = 0
quantidadeDeTokens = len(distribuicaoFrequencia.keys())
margemParaDesconsideracao = 0.15 * quantidadeDeTokens
limiteInferior = margemParaDesconsideracao
limiteSuperior = quantidadeDeTokens - margemParaDesconsideracao

for tokenFrequencia in distribuicaoFrequencia.keys():
    if (contador > limiteInferior) and (contador < limiteSuperior):
        if distribuicaoFrequencia[tokenFrequencia] == 0:
            tfidf = 0
        else:
            tfidf = (1 +
log(distribuicaoFrequencia[tokenFrequencia])) *
log(2285/distribuicaoFrequencia[tokenFrequencia])
            print 'Frequencia de tokens: ' + tokenFrequencia + ' -> ' +
tfidf
                listaPreliminarDeDescritores.append(tokenFrequencia)
                contador = contador + 1

#####
# Extracao definitiva de descritores: analise semantica
#####

for conceitoOntologia in listaDeConceitosOntologia:

```

```

        for descritorPreliminar in listaPreliminarDeDescritores:
            if descritorPreliminar ==
conceitoOntologia.attributes['rdf:ID'].value:
                print 'Descritor definitivo: ' + descritorPreliminar
                listaDefinitivaDeDescritores.append(descritorPreliminar)
                for propriedadeDoConceitoOntologia in
conceitoOntologia.childNodes:
                    if propriedadeDoConceitoOntologia.nodeType == 1:
                        if propriedadeDoConceitoOntologia.tagName ==
'Sinonimo':
                            print 'Descritor definitivo sinonimo:
' + propriedadeDoConceitoOntologia.attributes.item(0).value
                                listaDefinitivaDeDescritores.append(propriedadeDoConceitoOntologia.attributes
.item(0).value)
                                    elif propriedadeDoConceitoOntologia.tagName
== 'Holonimo':
                                        print 'Descritor definitivo holonimo:
' + propriedadeDoConceitoOntologia.attributes.item(0).value
                                            listaDefinitivaDeDescritores.append(propriedadeDoConceitoOntologia.attributes
.item(0).value)
                                                elif propriedadeDoConceitoOntologia.tagName
== 'Hiperonimo':
                                                    print 'Descritor definitivo
Hiperonimo: ' + propriedadeDoConceitoOntologia.attributes.item(0).value
                                                        listaDefinitivaDeDescritores.append(propriedadeDoConceitoOntologia.attributes
.item(0).value)
                                                            listaDetokens = listaDeLaudos.words(fileids=laudo)
                                                            for token in listaDetokens:
                                                                if token == conceitoOntologia.attributes['rdf:ID'].value:
                                                                    for propriedadeDoConceitoOntologia in
conceitoOntologia.childNodes:
                                                                        if propriedadeDoConceitoOntologia.nodeType == 1:
                                                                            if propriedadeDoConceitoOntologia.tagName ==
'Sinonimo':
                                                                                if
propriedadeDoConceitoOntologia.attributes.item(0).value == 'TRUE':
                                                                                    print 'Descritor definitivo com
propriedade Indice: ' + token
                                                                                        listaDefinitivaDeDescritores.append(token)
                                                                                            #####
                                                                                            # Gravacao do arquivo de indexacao automatica
                                                                                            #
                                                                                            #####
nomeArquivoIA = diretorioLaudos + '/' + str(laudo)[-4] + ' IA.txt'
arquivoIA = open(nomeArquivoIA, 'w')
for descritorDefinitivo in listaDefinitivaDeDescritores:
    arquivoIA.write(descritorDefinitivo + '\r\n')
arquivoIA.close()

#####
# Conclusao
#####

print '!!! Indexacao automatica concluida com sucesso !!!'

#####
# Fim do script
#####

```

APÊNDICE B – ONTOLOGIA

```
; Thu Apr 11 14:30:00 BRT 2013
;
;+ (version "3.4")
;+ (build "Build 130")

(defclass %3ACLIPS_TOP_LEVEL_SLOT_CLASS "Fake class to save top-level slot
information"
  (is-a USER)
  (role abstract)
  (single-slot Meio
    (type INSTANCE)
;+    (allowed-classes Relacao)
;+    (cardinality 0 1)
    (create-accessor read-write))
  (single-slot Definicao
    (type STRING)
;+    (cardinality 0 1)
    (create-accessor read-write))
  (single-slot Lugar
    (type INSTANCE)
;+    (allowed-classes Relacao)
;+    (cardinality 0 1)
    (create-accessor read-write))
  (multislot Hiperonimo
    (type INSTANCE)
;+    (allowed-classes Conceito)
    (create-accessor read-write))
  (multislot Holonimo
    (type INSTANCE)
;+    (allowed-classes Conceito)
    (create-accessor read-write))
  (multislot Nao
    (type INSTANCE)
;+    (allowed-classes Conceito)
    (create-accessor read-write))
  (multislot Sem
    (type INSTANCE)
;+    (allowed-classes Conceito)
    (create-accessor read-write))
  (single-slot Instrumento
    (type INSTANCE)
;+    (allowed-classes Relacao))
```

```

;+          (cardinality 0 1)
            (create-accessor read-write))
(single-slot Inv
  (type INSTANCE)
;+          (allowed-classes Conceito)
;+          (cardinality 0 1)
            (create-accessor read-write))
(single-slot Valor
  (type STRING)
;+          (cardinality 0 1)
            (create-accessor read-write))
(single-slot Agente
  (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
            (create-accessor read-write))
(multislot RelaxavelA
  (type INSTANCE)
;+          (allowed-classes Conceito)
            (create-accessor read-write))
(single-slot Tema
  (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
            (create-accessor read-write))
(single-slot MedidaPadrao
  (type STRING)
;+          (cardinality 0 1)
            (create-accessor read-write))
(single-slot Indice
  (type SYMBOL)
  (allowed-values FALSE TRUE)
;+          (cardinality 0 1)
            (create-accessor read-write))
(single-slot Paciente
  (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
            (create-accessor read-write))
(single-slot Destino
  (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
;+          (inverse-slot Fonte)
            (create-accessor read-write))
(single-slot Fonte

```

```

                (type INSTANCE)
;+            (allowed-classes Relacao)
;+            (cardinality 0 1)
;+            (inverse-slot Destino)
                (create-accessor read-write))
(multislot Eum
                (type INSTANCE)
;+            (allowed-classes Conceito)
                (create-accessor read-write))
(single-slot Padrao
                (type INSTANCE)
;+            (allowed-classes Conceito)
;+            (cardinality 0 1)
                (create-accessor read-write))
(single-slot Rota
                (type INSTANCE)
;+            (allowed-classes Relacao)
;+            (cardinality 0 1)
                (create-accessor read-write))
(multislot Sinonimo
                (type INSTANCE)
;+            (allowed-classes Conceito)
                (create-accessor read-write))
(multislot Origem
                (type INSTANCE)
;+            (allowed-classes Conceito)
                (create-accessor read-write))
(multislot EspacoTemporal
                (type INSTANCE)
;+            (allowed-classes Conceito)
                (create-accessor read-write)))

(defclass Conceito
  (is-a USER)
  (role concrete)
  (single-slot Meio
    (type INSTANCE)
;+    (allowed-classes Relacao)
;+    (cardinality 0 1)
    (create-accessor read-write))
  (single-slot Definicao
    (type STRING)
;+    (cardinality 0 1)
    (create-accessor read-write))
  (single-slot Lugar
    (type INSTANCE)

```

```

;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
           (create-accessor read-write))
(multislot Hiperonimo
           (type INSTANCE)
;+          (allowed-classes Conceito)
           (create-accessor read-write))
(multislot Holonimo
           (type INSTANCE)
;+          (allowed-classes Conceito)
           (create-accessor read-write))
(single-slot Instrumento
           (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
           (create-accessor read-write))
(single-slot Agente
           (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
           (create-accessor read-write))
(single-slot Tema
           (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
           (create-accessor read-write))
(single-slot Indice
           (type SYMBOL)
           (allowed-values FALSE TRUE)
;+          (cardinality 0 1)
           (create-accessor read-write))
(single-slot Paciente
           (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
           (create-accessor read-write))
(single-slot Destino
           (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
           (create-accessor read-write))
(single-slot Fonte
           (type INSTANCE)
;+          (allowed-classes Relacao)
;+          (cardinality 0 1)
           (create-accessor read-write))

```



```

(multislot Eum
  (type INSTANCE)
;+   (allowed-classes Conceito)
      (create-accessor read-write))
(single-slot Rota
  (type INSTANCE)
;+   (allowed-classes Relacao)
;+   (cardinality 0 1)
      (create-accessor read-write))
(multislot Sinonimo
  (type INSTANCE)
;+   (allowed-classes Conceito)
      (create-accessor read-write)))

(defclass Relacao
  (is-a USER)
  (role concrete)
  (single-slot Inv
    (type INSTANCE)
;+   (allowed-classes Conceito)
;+   (cardinality 0 1)
      (create-accessor read-write))
  (single-slot Valor
    (type STRING)
;+   (cardinality 0 1)
      (create-accessor read-write))
  (multislot RelaxavelA
    (type INSTANCE)
;+   (allowed-classes Conceito)
      (create-accessor read-write))
  (single-slot MedidaPadrao
    (type STRING)
;+   (cardinality 0 1)
      (create-accessor read-write))
  (multislot Nao
    (type INSTANCE)
;+   (allowed-classes Conceito)
      (create-accessor read-write))
  (single-slot Padrao
    (type INSTANCE)
;+   (allowed-classes Conceito)
;+   (cardinality 0 1)
      (create-accessor read-write))
  (multislot Sem
    (type INSTANCE)
;+   (allowed-classes Conceito)

```

```

                (create-accessor read-write))
        (multislot Origem
            (type INSTANCE)
            (allowed-classes Conceito)
            (create-accessor read-write))
;+
        (multislot EspacoTemporal
            (type INSTANCE)
            (allowed-classes Conceito)
            (create-accessor read-write)))

<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns="http://www.owl-ontologies.com/unnamed.owl#"
    xml:base="http://www.owl-ontologies.com/unnamed.owl">
    <owl:Ontology rdf:about=""/>
    <owl:Class rdf:ID="Conceito"/>
    <owl:Class rdf:ID="Relacao"/>
    <owl:ObjectProperty rdf:ID="Sinonimo">
        <rdfs:domain rdf:resource="#Conceito"/>
        <rdfs:range rdf:resource="#Conceito"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="Hiperonimo">
        <rdfs:domain rdf:resource="#Conceito"/>
        <rdfs:range rdf:resource="#Conceito"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="Nao">
        <rdfs:range rdf:resource="#Conceito"/>
        <rdfs:domain rdf:resource="#Relacao"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="Holonimo">
        <rdfs:domain rdf:resource="#Conceito"/>
        <rdfs:range rdf:resource="#Conceito"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="RelaxavelA">
        <rdfs:domain rdf:resource="#Relacao"/>
        <rdfs:range rdf:resource="#Conceito"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="Eum">
        <rdfs:domain rdf:resource="#Conceito"/>
        <rdfs:range rdf:resource="#Conceito"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="EspacoTemporal">
        <rdfs:range rdf:resource="#Conceito"/>
        <rdfs:domain rdf:resource="#Relacao"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="Origem">
        <rdfs:range rdf:resource="#Conceito"/>
        <rdfs:domain rdf:resource="#Relacao"/>
    </owl:ObjectProperty>
    <owl:ObjectProperty rdf:ID="Sem">
        <rdfs:range rdf:resource="#Conceito"/>
        <rdfs:domain rdf:resource="#Relacao"/>
    </owl:ObjectProperty>
    <owl:FunctionalProperty rdf:ID="Inv">
        <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
        <rdfs:domain rdf:resource="#Relacao"/>
        <rdfs:range rdf:resource="#Conceito"/>
    </owl:FunctionalProperty>
    <owl:FunctionalProperty rdf:ID="Agente">

```

```

    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <rdfs:domain rdf:resource="#Conceito"/>
    <rdfs:range rdf:resource="#Relacao"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Meio">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <rdfs:domain rdf:resource="#Conceito"/>
    <rdfs:range rdf:resource="#Relacao"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Definicao">
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
    <rdfs:domain rdf:resource="#Conceito"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Valor">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:domain rdf:resource="#Relacao"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="MedidaPadrao">
    <rdfs:domain rdf:resource="#Relacao"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Tema">
    <rdfs:range rdf:resource="#Relacao"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <rdfs:domain rdf:resource="#Conceito"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Destino">
    <rdfs:range rdf:resource="#Relacao"/>
    <rdfs:domain rdf:resource="#Conceito"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <owl:inverseOf>
        <owl:FunctionalProperty rdf:ID="Fonte"/>
    </owl:inverseOf>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Padrao">
    <rdfs:range rdf:resource="#Conceito"/>
    <rdfs:domain rdf:resource="#Relacao"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Instrumento">
    <rdfs:range rdf:resource="#Relacao"/>
    <rdfs:domain rdf:resource="#Conceito"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Paciente">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <rdfs:range rdf:resource="#Relacao"/>
    <rdfs:domain rdf:resource="#Conceito"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Lugar">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <rdfs:range rdf:resource="#Relacao"/>
    <rdfs:domain rdf:resource="#Conceito"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Rota">
    <rdfs:range rdf:resource="#Relacao"/>
    <rdfs:domain rdf:resource="#Conceito"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="Indice">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
    <rdfs:domain rdf:resource="#Conceito"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:about="#Fonte">

```

```

<owl:inverseOf rdf:resource="#Destino"/>
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
<rdfs:range rdf:resource="#Relacao"/>
<rdfs:domain rdf:resource="#Conceito"/>
</owl:FunctionalProperty>
<Conceito rdf:ID="FRAUDULENTA_UTILIZACAO_DE_RECIPIENTE"/>
<Conceito rdf:ID="COMPRA_DE_VOTO">
  <Hiperonimo>
    <Conceito rdf:ID="COMPRA"/>
  </Hiperonimo>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">
    true</Indice>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    COMPRA DE VOTO EM ELEICAO</Definicao>
</Conceito>
<Conceito rdf:ID="PERICIA_GRAFOTECNICA">
  <Enum>
    <Conceito rdf:ID="PERICIA">
      <Agente>
        <Relacao rdf:ID="RelacaoOntologica_Agente_PERICIA_9">
          <Padrao>
            <Conceito rdf:ID="PERITO_CRIMINAL">
              <Holonimo>
                <Conceito rdf:ID="PERITO_OFICIAL">
                  <Instrumento>
                    <Relacao
rdf:ID="RelacaoOntologica_Instrumento_PERITO_OFICIAL_1">
                      <Sem rdf:resource="#PERICIA"/>
                    </Relacao>
                  </Instrumento>
                </Hiperonimo>
                <Conceito rdf:ID="PERITO"/>
              </Hiperonimo>
              <Holonimo rdf:resource="#PERITO_CRIMINAL"/>
            </Conceito>
          </Holonimo>
          <Hiperonimo rdf:resource="#PERITO"/>
        </Conceito>
      </Padrao>
      <RelaxavelA rdf:resource="#PERITO_OFICIAL"/>
    </Relacao>
  </Agente>
</Conceito>
</Enum>
</Conceito>
<Conceito rdf:ID="FRAUDE_PROCESSUAL"/>
<Conceito rdf:ID="FRAUDE_DE_PRECO"/>
<Conceito rdf:ID="PERITO_NAO-OFICIAL">
  <Hiperonimo rdf:resource="#PERITO"/>
  <Holonimo rdf:resource="#PERITO_OFICIAL"/>
</Conceito>
<Conceito rdf:ID="CARTAO_DE_CREDITO"/>
<Conceito rdf:ID="INSTITUICAO_BANCARIA_INTERNACIONAL">
  <Hiperonimo>
    <Conceito rdf:ID="BANCO_INTERNACIONAL">
      <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        BANCO INTERNACIONAL</Definicao>
    </Conceito>
  </Hiperonimo>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    INSTITUICAO BANCARIA INTERNACIONAL</Definicao>
</Conceito>
<Conceito rdf:ID="COPIA_ILEGIVEL">
  <Holonimo>
    <Conceito rdf:ID="COPIA">
      <Enum rdf:resource="#COPIA_ILEGIVEL"/>
    </Conceito>
  </Holonimo>

```

```

</Conceito>
<Conceito rdf:ID="SEXO_COM_ADOLESCENTE">
  <Eum>
    <Conceito rdf:ID="SEXO">
      <Holonimo>
        <Conceito rdf:ID="PEDOFILIA">
          <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
            >PEDOFILIA</Definicao>
          <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">
            >true</Indice>
        </Conceito>
      </Holonimo>
      <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        >SEXO</Definicao>
    </Conceito>
  </Eum>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">
    >true</Indice>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    >Sexo com Adolescente</Definicao>
</Conceito>
<Conceito rdf:ID="NEGACAO"/>
<Conceito rdf:ID="INJURIA_REAL">
  <Hiperonimo>
    <Conceito rdf:ID="INJURIA">
      <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">
        >true</Indice>
      <Holonimo>
        <Conceito rdf:ID="DIFAMACAO">
          <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">
            >true</Indice>
          <Holonimo rdf:resource="#INJURIA"/>
        </Conceito>
      </Holonimo>
    </Conceito>
  </Hiperonimo>
</Eum>
  <Conceito rdf:ID="INJURIA_GRAVE">
    <Hiperonimo rdf:resource="#INJURIA"/>
    <Eum rdf:resource="#INJURIA_REAL"/>
  </Conceito>
</Eum>
</Conceito>
<Conceito rdf:ID="DISSEMINACAO_DE_PROGRAMA_MALICIOSO">
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">
    >true</Indice>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    >DISSEMINACAO DE PROGRAMA DE COMPUTADOR COM FINS MALICIOSOS</Definicao>
</Conceito>
<Conceito rdf:ID="INSTITUICAO_FINANCEIRA_PUBLICA">
  <Eum>
    <Conceito rdf:ID="INSTITUICAO_FINANCEIRA_PRIVADA">
      <Eum rdf:resource="#INSTITUICAO_FINANCEIRA_PUBLICA"/>
      <Hiperonimo>
        <Conceito rdf:ID="INSTITUICAO_FINANCEIRA">
          <Holonimo>
            <Conceito rdf:ID="CONTABILIDADE">
              <Holonimo rdf:resource="#INSTITUICAO_FINANCEIRA"/>
            </Conceito>
          </Holonimo>
        </Conceito>
      </Hiperonimo>
    </Conceito>
  </Eum>
  <Hiperonimo rdf:resource="#INSTITUICAO_FINANCEIRA"/>
</Conceito>
<Conceito rdf:ID="FRAUDE_EM_PARECER">
  <Eum>

```

```

<Conceito rdf:ID="FRAUDE_EM_INFORMACAO">
  <Holonimo rdf:resource="#FRAUDE_EM_PARECER"/>
  <Holonimo>
    <Conceito rdf:ID="FRAUDE_EM_RELATORIO">
      <Holonimo rdf:resource="#FRAUDE_EM_PARECER"/>
      <Eum rdf:resource="#FRAUDE_EM_INFORMACAO"/>
      <Eum>
        <Conceito rdf:ID="FRAUDE_EM_LANCAMENTO">
          <Holonimo rdf:resource="#FRAUDE_EM_RELATORIO"/>
          <Holonimo>
            <Conceito rdf:ID="FRAUDE_EM_ESCRITURACAO">
              <Eum>
                <Conceito rdf:ID="FRAUDE_EM_REGISTRO">
                  <Holonimo rdf:resource="#FRAUDE_EM_LANCAMENTO"/>
                  <Eum rdf:resource="#FRAUDE_EM_ESCRITURACAO"/>
                </Conceito>
              </Eum>
            <Holonimo rdf:resource="#FRAUDE_EM_LANCAMENTO"/>
          </Conceito>
        </Holonimo>
      <Holonimo rdf:resource="#FRAUDE_EM_REGISTRO"/>
    </Conceito>
  </Eum>
</Conceito>
</Holonimo>
</Conceito>
</Eum>
<Eum rdf:resource="#FRAUDE_EM_RELATORIO"/>
</Conceito>
<Conceito rdf:ID="PERICIA_CONTABIL">
  <Eum rdf:resource="#PERICIA"/>
</Conceito>
<Conceito rdf:ID="FRAUDE_PARA_RECEBIMENTO_DE_SEGURO">
  <Eum>
    <Conceito rdf:ID="FRAUDE_PARA_RECEBIMENTO_DE_INDENIZACAO">
      <Holonimo rdf:resource="#FRAUDE_PARA_RECEBIMENTO_DE_SEGURO"/>
    </Conceito>
  </Eum>
</Conceito>
<Conceito rdf:ID="FRAUDE_NO_COMERCIO">
  <Paciente>
    <Relacao rdf:ID="RelacaoOntologica_Paciente_FRAUDE_NO_COMERCIO_3">
      <Sem>
        <Conceito rdf:ID="FRAUDE_NA_ENTREGA_DE_COISA">
          <Holonimo>
            <Conceito rdf:ID="VANTAGEM_INDEVIDA">
              <Eum>
                <Conceito rdf:ID="CORRUPCAO_PASSIVA">
                  <Eum rdf:resource="#VANTAGEM_INDEVIDA"/>
                <Holonimo>
                  <Conceito rdf:ID="CORRUPCAO_ATIVA">
                    <Indice
rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
                    >true</Indice>
                  <Eum rdf:resource="#VANTAGEM_INDEVIDA"/>
                  <Eum rdf:resource="#CORRUPCAO_PASSIVA"/>
                </Conceito>
              </Holonimo>
            </Conceito>
          </Eum>
        <Meio>
          <Relacao rdf:ID="RelacaoOntologica_Meio_VANTAGEM_INDEVIDA_8">
            <Sem rdf:resource="#FRAUDE_NA_ENTREGA_DE_COISA"/>
          </Relacao>
        </Meio>
      <Holonimo>
        <Conceito rdf:ID="VANTAGEM_ILICITA">
          <Holonimo rdf:resource="#VANTAGEM_INDEVIDA"/>
        </Conceito>
      </Holonimo>
    </Relacao>
  </Paciente>
</Conceito>

```

```

        </Conceito>
        </Holonimo>
        <Holonimo rdf:resource="#CORRUPCAO_ATIVA"/>
        </Conceito>
        </Holonimo>
        <Holonimo rdf:resource="#FRAUDE_NO_COMERCIO"/>
        </Conceito>
    </Sem>
</Relacao>
</Paciente>
</Conceito>
<Relacao rdf:ID="RelacaoOntologica_Agente_LAUDO_TECNICO_0">
    <Padrao rdf:resource="#PERITO_CRIMINAL"/>
</Relacao>
<Conceito rdf:ID="ELEICAO">
    <Holonimo>
        <Conceito rdf:ID="SUFRAGIO_UNIVERSAL">
            <Holonimo rdf:resource="#ELEICAO"/>
            <Eum>
                <Conceito rdf:ID="SUFRAGIO">
                    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
                    >SUFRAGIO</Definicao>
                </Conceito>
            </Eum>
        </Conceito>
    </Holonimo>
</Conceito>
<Conceito rdf:ID="SONEGACAO"/>
<Conceito rdf:ID="SEXO_COM_CRIANCA">
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Sexo com Criança</Definicao>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
    <Eum rdf:resource="#SEXO"/>
</Conceito>
<Conceito rdf:ID="NEGACAO_DE_SERVICO">
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >ATAQUE DE NEGACAO DE SERVICO 'DOS'</Definicao>
    <Hiperonimo rdf:resource="#NEGACAO"/>
</Conceito>
<Conceito rdf:ID="CONTABILIDADE_DE_JOGO">
    <Eum rdf:resource="#CONTABILIDADE"/>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >CONTABILIDADE DE JOGO</Definicao>
</Conceito>
<Conceito rdf:ID="PERICIA_MEDICA">
    <Eum rdf:resource="#PERICIA"/>
</Conceito>
<Conceito rdf:ID="FRAUDE_PREVIDENCIARIA">
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >FRAUDE PREVIDENCIARIA</Definicao>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
    <Eum>
        <Conceito rdf:ID="FRAUDE">
            <Holonimo>
                <Conceito rdf:ID="FRAUDE_DE_LEI_SOBRE_ESTRANGEIRO">
                    <Eum rdf:resource="#FRAUDE"/>
                </Conceito>
            </Holonimo>
        </Conceito>
    </Eum>
</Conceito>
<Conceito rdf:ID="CORRUPCAO_DE_MENORES">
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>

```

```

</Conceito>
<Conceito rdf:ID="CACA_NIQUEL">
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
  <Hiperonimo>
    <Conceito rdf:ID="MAQUINA_ELETRONICA">
      <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >MAQUINA ELETRONICA DE JOGO DE AZAR</Definicao>
      <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
      >true</Indice>
      <Eum>
        <Conceito rdf:ID="MAQUINA"/>
      </Eum>
    </Conceito>
  </Hiperonimo>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >MAQUINA CACA-NIQUEL</Definicao>
</Conceito>
<Conceito rdf:ID="FRAUDE_EM_ARREMATACAO_JUDICIAL"/>
<Conceito rdf:ID="INSS"/>
<Conceito rdf:ID="INSTITUICAO_FINANCEIRA_EXTRANGEIRA">
  <Hiperonimo rdf:resource="#BANCO_INTERNACIONAL"/>
  <Hiperonimo rdf:resource="#INSTITUICAO_FINANCEIRA"/>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >INSTITUICAO FINANCEIRA EXTRANGEIRA</Definicao>
</Conceito>
<Relacao rdf:ID="RelacaoOntologica_Lugar_JOGO_DE_AZAR_0">
  <Padrao>
    <Conceito rdf:ID="CASSINO">
      <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >LOCAL PARA PRATICA DE JOGO DE AZAR</Definicao>
      <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
      >true</Indice>
    </Conceito>
  </Padrao>
</Relacao>
<Conceito rdf:ID="PROJETO_BASICO">
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >PROJETO BASICO</Definicao>
  <Hiperonimo>
    <Conceito rdf:ID="TERMO_DE_REFERENCIA">
      <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >TERMO DE REFERENCIA</Definicao>
    </Conceito>
  </Hiperonimo>
</Conceito>
<Conceito rdf:ID="IMAGEM_PORNOGRAFICA_DE_ADOLESCENTE">
  <Eum>
    <Conceito rdf:ID="PORNOGRAFIA"/>
  </Eum>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Imagem Pornogrã;fica de Adolescente</Definicao>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
  <Hiperonimo rdf:resource="#PEDOFILIA"/>
</Conceito>
<Conceito rdf:ID="CORRUPCAO">
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
</Conceito>
<Relacao rdf:ID="RelacaoOntologica_Meio_FRAUDE_BANCARIA_0">
  <Sem>
    <Conceito rdf:ID="FRAUDE_ELETRONICA">
      <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
      >true</Indice>
    </Conceito>
  </Sem>
</Relacao>

```



```

<Conceito rdf:ID="COPIA_ILEGAL">
  <Sinonimo>
    <Conceito rdf:ID="COPIA_PIRATA">
      <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >COPIA PIRATA</Definicao>
      <Holonimo>
        <Conceito rdf:ID="PIRATARIA">
          <Hiperonimo>
            <Conceito rdf:ID="VIOLACAO_DE_DIREITO_AUTORAL">
              <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
              >true</Indice>
            </Conceito>
          </Hiperonimo>
          <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >PIRATARIA</Definicao>
        </Conceito>
      </Holonimo>
      <Eum rdf:resource="#COPIA"/>
    </Conceito>
  </Sinonimo>
  <Eum rdf:resource="#COPIA"/>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >COPIA ILEGAL</Definicao>
</Conceito>
<Conceito rdf:ID="SONEGACAO_FISCAL">
  <Hiperonimo rdf:resource="#SONEGACAO"/>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
  <Eum rdf:resource="#SONEGACAO"/>
  <Holonimo>
    <Conceito rdf:ID="FRAUDE_FISCAL">
      <Holonimo rdf:resource="#SONEGACAO_FISCAL"/>
    </Conceito>
  </Holonimo>
</Conceito>
<Conceito rdf:ID="FRAUDE_NA_ADMINISTRACAO_DE_SOCIEDADE_POR_ACOES">
  <Holonimo>
    <Conceito rdf:ID="FRAUDE_NA_FUNDACAO_DE_SOCIEDADE_POR_ACOES">
      <Holonimo rdf:resource="#FRAUDE_NA_ADMINISTRACAO_DE_SOCIEDADE_POR_ACOES"/>
    </Conceito>
  </Holonimo>
</Conceito>
<Conceito rdf:ID="LAUDO_TECNICO">
  <Agente rdf:resource="#RelacaoOntologica_Agente_LAUDO_TECNICO_0"/>
</Conceito>
<Conceito rdf:ID="CONGESTIONAMENTO_DE_SERVICO">
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >CONGESTIONAMENTO DE SERVICO</Definicao>
</Conceito>
<Conceito rdf:ID="TRAFICO_INTERNACIONAL_DE_TOXICOS">
  <Sinonimo>
    <Conceito rdf:ID="TRAFICO_INTERNACIONAL_DE_ENTORPECENTES">
      <Holonimo>
        <Conceito rdf:ID="TRAFICANTE">
          <Holonimo>
            <Conceito rdf:ID="TRAFICO_DE_ENTORPECENTES">
              <Eum rdf:resource="#TRAFICO_INTERNACIONAL_DE_ENTORPECENTES"/>
              <Eum rdf:resource="#TRAFICANTE"/>
              <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
              >true</Indice>
            <Agente>
              <Relacao
rdf:ID="RelacaoOntologica_Agente_TRAFICO_DE_ENTORPECENTES_2">
              <RelaxavelA>
                <Conceito rdf:ID="LAVAGEM_DE_DINHEIRO">
                  <Indice
rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
                  >true</Indice>

```

```

        <Definicao
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >ATO DE OCULTAR OU DISSIMULAR A NATUREZA, ORIGEM,
LOCALIZACAO, DISPOSICAO OU PROPRIEDADE DE BENS, DIREITOS OU VALORES
PROVENIENTES DIRETA OU INDIRETAMENTE DE CRIME.</Definicao>
        <Holonimo rdf:resource="#TRAFICO_DE_ENTORPECENTES"/>
        </Conceito>
        </RelaxavelA>
        </Relacao>
        </Agente>
        </Conceito>
        </Holonimo>
        <Eum rdf:resource="#TRAFICO_INTERNACIONAL_DE_ENTORPECENTES"/>
        </Conceito>
        </Holonimo>
        <Eum rdf:resource="#TRAFICO_DE_ENTORPECENTES"/>
        <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
        >true</Indice>
        <Hiperonimo>
        <Conceito rdf:ID="TRAFICO_INTERNACIONAL"/>
        </Hiperonimo>
        </Conceito>
        </Sinonimo>
        <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
        >true</Indice>
</Conceito>
<Conceito rdf:ID="PAGAMENTO_DE_CONTA">
        <Hiperonimo>
        <Conceito rdf:ID="FRAUDE_BANCARIA">
        <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >FRAUDE EM BANCO</Definicao>
        <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
        >true</Indice>
        <Eum rdf:resource="#FRAUDE"/>
        <Meio rdf:resource="#RelacaoOntologica_Meio_FRAUDE_BANCARIA_0"/>
        </Conceito>
        </Hiperonimo>
        <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
        >false</Indice>
        <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >PAGAMENTO DE CONTA DE TERCEIROS</Definicao>
        </Conceito>
        <Conceito rdf:ID="JOGO_DE_AZAR_CONTRAVENCAO"/>
        <Conceito rdf:ID="FRAUDE_DE_EXECUCAO"/>
        <Conceito rdf:ID="DROGA_DE_ABUSO">
        <Sinonimo>
        <Conceito rdf:ID="DROGA"/>
        </Sinonimo>
        <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >DROGA DE ABUSO</Definicao>
        </Conceito>
        <Conceito rdf:ID="LAVAGEM_DE_CAPITAIS">
        <Sinonimo rdf:resource="#LAVAGEM_DE_DINHEIRO"/>
        </Conceito>
        <Conceito rdf:ID="BOTNET">
        <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
        >true</Indice>
        <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >REDE ZUMBI PARA ATAQUE DE NEGACAO DE SERVICO 'DDOS'</Definicao>
        </Conceito>
        <Conceito rdf:ID="FRAUDE_DE_CONCORRENCIA"/>
        <Conceito rdf:ID="CONGESTIONAMENTO_DE_TRAFEGO">
        <Sinonimo>
        <Conceito rdf:ID="CONGESTIONAMENTO_DE_SERVICO_DE_REDE">
        <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >CONGESTIONAMENTO DE SERVICO DE REDE</Definicao>
        <Hiperonimo rdf:resource="#CONGESTIONAMENTO_DE_SERVICO"/>
        </Conceito>

```

```

    </Sinonimo>
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >CONGESTIONAMENTO DE TRAFEGO DE REDE</Definicao>
</Conceito>
<Conceito rdf:ID="INVASAO_DE_DISPOSITIVO_INFORMATICO">
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >INVASAO DE DISPOSITIVO INFORMATICO</Definicao>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
</Conceito>
<Conceito rdf:ID="FRAUDE_DE_PESO">
    <Enum>
        <Conceito rdf:ID="FRAUDE_DE_MEDIDA">
            <Holonimo rdf:resource="#FRAUDE_DE_PESO"/>
        </Conceito>
    </Enum>
</Conceito>
<Conceito rdf:ID="FRAUDE_CONTRA_CREDORES"/>
<Conceito rdf:ID="INSTITUICAO_BANCARIA_AMERICANA">
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >INSTITUICAO BANCARIA AMERICANA</Definicao>
    <Hiperonimo rdf:resource="#BANCO_INTERNACIONAL"/>
</Conceito>
<Conceito rdf:ID="ENRIQUECIMENTO_ILICITO">
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
    <Sinonimo>
        <Conceito rdf:ID="ENRIQUECIMENTO_SEM_CAUSA"/>
    </Sinonimo>
</Conceito>
<Conceito rdf:ID="SOFTWARE"/>
<Conceito rdf:ID="FRAUDULENTA_UTILIZACAO_DE_INVOLUCRO"/>
<Conceito rdf:ID="CONGESTIONAMENTO_DE_SERVIDOR">
    <Hiperonimo rdf:resource="#NEGACAO_DE_SERVICO"/>
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >CONGESTIONAMENTO DE SERVIDOR</Definicao>
</Conceito>
<Conceito rdf:ID="PROGRAMA_DE_JOGO">
    <Enum rdf:resource="#SOFTWARE"/>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >PROGRAMA DE JOGO DE AZAR</Definicao>
</Conceito>
<Conceito rdf:ID="EVASAO_MEDIANTE_VIOLENCIA">
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
</Conceito>
<Conceito rdf:ID="TRAFICO_INTERNACIONAL_DE_DROGAS">
    <Sinonimo rdf:resource="#TRAFICO_INTERNACIONAL_DE_ENTORPECENTES"/>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
</Conceito>
<Conceito rdf:ID="INVASAO_DE_REDE_DE_COMPUTADOR">
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >INVASAO DE REDE DE COMPUTADOR</Definicao>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
    <Enum rdf:resource="#INVASAO_DE_DISPOSITIVO_INFORMATICO"/>
</Conceito>
<Conceito rdf:ID="REGISTRO_DE_ATIVIDADE_SEXUAL_COM_CRIANCA">
    <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Registro de Atividade Sexual com Criança</Definicao>
    <Hiperonimo rdf:resource="#PEDOFILIA"/>
    <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</Indice>
    <Enum rdf:resource="#PORNOGRAFIA"/>
</Conceito>

```

```

<Conceito rdf:ID="FRAUDE_EM_LICITACAO">
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >FRAUDE EM PROCESSO LICITATORIO</Definicao>
  <Eum rdf:resource="#FRAUDE"/>
</Conceito>
<Conceito rdf:ID="JOGO_DE_AZAR">
  <Sinonimo rdf:resource="#JOGO_DE_AZAR_CONTRAVENCAO"/>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
  <Lugar rdf:resource="#RelacaoOntologica_Lugar_JOGO_DE_AZAR_0"/>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >JOGO DE AZAR</Definicao>
</Conceito>
<Conceito rdf:ID="TRAFICO_DE_INFLUENCIA">
  <Holonimo rdf:resource="#VANTAGEM_INDEVIDA"/>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >CRIME PREVISTO NO ARTIGO 332 DO CODIGO PENAL, COM REDACAO DA LEI Nº 9.127,
  DE 16/11/95.</Definicao>
</Conceito>
<Conceito rdf:ID="FRAUDE_NO_PAGAMENTO_POR_MEIO_DE_CHEQUE"/>
<Conceito rdf:ID="VANTAGEM"/>
<Conceito rdf:ID="VIRUS_DE_COMPUTADOR">
  <Eum rdf:resource="#SOFTWARE"/>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >VIRUS DE COMPUTADOR</Definicao>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
</Conceito>
<Conceito rdf:ID="PORNOGRAFIA_INFANTIL">
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >PORNOGRAFIA INFANTIL</Definicao>
  <Holonimo rdf:resource="#PEDOFILIA"/>
  <Eum rdf:resource="#PORNOGRAFIA"/>
</Conceito>
<Conceito rdf:ID="LAUDO_DE_AVALIACAO">
  <Hiperonimo rdf:resource="#LAUDO_TECNICO"/>
</Conceito>
<Conceito rdf:ID="SOFTWARE_SEM_LICENCA">
  <Sinonimo rdf:resource="#COPIA_PIRATA"/>
  <Eum rdf:resource="#SOFTWARE"/>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >SOFTWARE SEM LICENCA</Definicao>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
</Conceito>
<Conceito rdf:ID="PERICIA_ANTROPOLOGICA">
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >APURACAO DO GRAU DE INTEGRACAO DO INDIO A COMUNIDADE.</Definicao>
  <Eum rdf:resource="#PERICIA"/>
</Conceito>
<Conceito rdf:ID="VISTORIA">
  <Eum rdf:resource="#PERICIA"/>
</Conceito>
<Conceito rdf:ID="TRAFICO_DE_ENTORPECENTE">
  <Sinonimo rdf:resource="#TRAFICO_DE_ENTORPECENTES"/>
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
</Conceito>
<Conceito rdf:ID="EVASAO_FISCAL">
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
</Conceito>
<Conceito rdf:ID="APLICATIVO_COM_LICENCIAMENTO_IRREGULAR">
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"

```

```

>APLICATIVO COM LICENCIAMENTO IRREGULAR</Definicao>
<Sinonimo rdf:resource="#COPIA_PIRATA"/>
<Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
>true</Indice>
<Eum rdf:resource="#SOFTWARE"/>
</Conceito>
<Conceito rdf:ID="INSTITUICAO_BANCARIA_EUROPEIA">
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >INSTITUICAO BANCARIA EUROPEIA</Definicao>
  <Hiperonimo rdf:resource="#BANCO_INTERNACIONAL"/>
</Conceito>
<Conceito rdf:ID="EVASAO_DE_DIVISAS">
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
</Conceito>
<Conceito rdf:ID="INVASAO_DE_COMPUTADOR_PESSOAL">
  <Indice rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
  >true</Indice>
  <Definicao rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >INVASAO DE COMPUTADOR PESSOAL</Definicao>
  <Eum rdf:resource="#INVASAO_DE_DISPOSITIVO_INFORMATICO"/>
</Conceito>
</rdf:RDF>

<!-- Created with Protege (with OWL Plugin 3.4, Build 130)
http://protege.stanford.edu -->

```