PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CONSTRUÇÃO DE ESTRUTURAS ONTOLÓGICAS A PARTIR DE TEXTOS: UM ESTUDO BASEADO NO MÉTODO FORMAL CONCEPT ANALYSIS E EM PAPÉIS SEMÂNTICOS

SÍLVIA MARIA WANDERLEY MORAES

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profa. Dra. Vera Lúcia Strube de Lima

Porto Alegre
2012

M827c Moraes, Sílvia Maria Wanderley
Construção de estruturas ontológicas a partir de textos : um estudo baseado no método formal concept analysis e em papéis semânticos / Sílvia Maria Wanderley Moraes. – Porto Alegre, 2012.
184 f.

Tese (Doutorado) – Fac. de Informática, PUCRS.
Orientador: Profª. Drª. Vera Lúcia Strube de Lima.

1. Informática. 2. Ontologia. 3. Análise Semântica (Programação). 4. Processamento da Linguagem Natural. I. Lima, Vera Lúcia Strube de. II. Título.
CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "Construção de Estruturas Ontológicas a Partir de Textos: Um Estudo Baseado no Método Formal *Concept Analysis* e em Papéis Semânticos", apresentada por Sílvia Maria Wanderley Moraes, como parte dos requisitos para obtenção do grau de Doutor em Ciência da Computação, Inteligência Computacional, aprovada em 30/03/2012 pela Comissão Examinadora:

Vera Lúcia Strube de Lima

Profa. Dra. Vera Lúcia Strube de Lima – Orientadora — PPGCC/PUCRS

Profa. Dra. Renata Vieira - — PPGCC/PUCRS

Prof. Dr. Antonio Carlos da Rocha Costa - — FURG

Prof. Dr. Ruy Luiz Milidiú - — PUC-Rio

Homologada em 22/06/2012, conforme Ata No. 013 pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

DEDICATÓRIA

Dedico esta tese de doutorado à minha família: à minha mãe, ao Diego e às minhas duas bençãos de Deus - meus filhos Vítor, de 2 anos, e Lucas, que ainda carrego comigo.

AGRADECIMENTOS

Agradeço a Deus, primeiramente, por conseguir concluir esta tese sem qualquer prejuízo ao Lucas. Agradeço muito ao Diego pelo companheirismo e pela grande compreensão demonstrada em relação a todos os momentos que estive ausente em decorrência deste trabalho. Seu incentivo para conclusão da tese foi muito importante. Meu agradecimento especial à minha mãe, Adélia, sem a qual eu não conseguiria levar adiante este projeto profissional. Enquanto eu trabalhava, ela cuidava com carinho sem igual da minha família. Também agradeço à tia Marli pela atenção que recebi todas as vezes que fui ao Rio de Janeiro por razões relacionadas a este trabalho. Agradeço igualmente à tia Martha, que com frequência vinha nos ajudar a cuidar do Vítor, e, desta forma, permitiu que eu usasse esses momentos para me dedicar à tese.

Agradeço à minha orientadora, Prof. Vera, pelas suas preciosas contribuições, disponibilidade e incentivo no decorrer deste trabalho. Agradeço também pela compreensão em relação aos momentos que tive que me afastar para cuidar do Vítor. Agradeço ao Prof. Ruy por gentilmente ter me recebido na PUC-Rio e permitido, na época, meu acesso ao processador de papéis de semânticos F-EXT-WS, o qual utilizo nesta tese. Agradeço também ao Prof. António Branco e a sua equipe de pesquisadores. Todos que me receberam muito bem e com muita atenção, quando estive em Lisboa, por questões relativas a um projeto conjunto entre o nosso grupo de pesquisa e o dele, durante o doutoramento. Agradeço também aos professores da banca por terem aceito a tarefa de avaliar meu trabalho.

Agradeço, ainda, ao meu amigo Marcelo Cohen pela sua amizade e inestimável ajuda. Ele sempre se dispôs a me auxiliar nos momentos em que eu precisava resolver problemas relativos à conversão de algum arquivo, configuração de ferramentas e dúvidas relacionadas ao Lyx e ao LaTeX. Agradeço também ao Marco Gonzalez por ter disponibilizado a ferramenta FORMA e resolvido dúvidas quanto ao seu uso. Agradeço ao Marco Mangan por ter me indicado um artigo que foi fundamental para a organização dos capítulos referentes à tese.

Agradeço aos meus amigos e colegas pelos momentos de descontração na sala dos professores horistas da PUCRS, os quais tornaram o caminho até a conclusão da tese mais ameno. São eles: Márcia Moraes, Letícia Leite, Cristina Nunes, Fernanda Walker, Tiago e Rodrigo Espíndola. Agradeço, por fim, a todas as pessoas que de alguma forma me ajudaram a concluir esta tese.

CONSTRUÇÃO DE ESTRUTURAS ONTOLÓGICAS A PARTIR DE TEXTOS: UM ESTUDO BASEADO NO MÉTODO FORMAL CONCEPT ANALYSIS E EM PAPÉIS SEMÂNTICOS

RESUMO

Este trabalho tem como propósito estudar estruturas conceituais geradas seguindo o método Formal Concept Analysis. Usamos na construção dessas estruturas informações lexicossemânticas extraídas dos textos, dentre as quais se destacam os papéis semânticos. Em nossa pesquisa, propomos formas de inclusão de tais papéis nos conceitos produzidos por esse método formal. Analisamos a contribuição dos papéis semânticos e das classes de verbos na composição dos conceitos, por meio de medidas de ordem estrutural. Nesses estudos, utilizamos os *corpora* Penn TreeBank Sample e SemLink 1.1, ambos em Língua Inglesa. Testamos, também para Língua Inglesa, a aplicabilidade de nossa proposta nos domínios de Finanças e Turismo com textos extraídos do *corpus* Wikicorpus 1.0. Essa aplicabilidade foi analisada extrinsecamente com base na tarefa de categorização de textos, a qual foi avaliada a partir de medidas de ordem funcional tradicionalmente usadas nessa área. Realizamos ainda alguns estudos preliminares relacionados à nossa proposta para um *corpus* em Língua Portuguesa: PLN-BR CATEG. Obtivemos, nos estudos realizados, resultados satisfatórios os quais mostram que a abordagem proposta é promissora.

Palavras-chave: Processamento de Linguagem Natural; Estruturas Conceituais; Relações Não Taxonômicas; Papéis Semânticos; Formal Concept Analysis.

ONTOLOGICAL STRUCTURES BUILDING FROM TEXTS: A STUDY BASED ON FORMAL CONCEPT ANALYSIS METHOD AND SEMANTIC ROLES

ABSTRACT

This work aims to study conceptual structures based on the Formal Concept Analysis method. We build these structures based on lexico-semantic information extracted from texts, among which we highlight the semantic roles. In our research, we propose ways to include semantic roles in concepts produced by this formal method. We analyze the contribution of semantic roles and verb classes in the composition of these concepts through structural measures. In these studies, we use the Penn Treebank Sample and SemLink 1.1 *corpora*, both in English. We test, also for English, the applicability of our proposal in the Finance and Tourism domains with text extracted from the Wikicorpus 1.0. This applicability was extrinsically analyzed based on the text categorization task, which was evaluated through functional measures traditionally used in this area. We also performed some preliminary studies for a *corpus* in Portuguese: PLN-BR CATEG. In our studies, we obtained satisfactory results which show that the proposed approach is promising.

Keywords: Natural Language Processing; Conceptual Structures; Non-Taxonomic Relations; Semantic Roles; Formal Concept Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AE	Above Expectation
AL	Abrangência Léxica
AT	Abrangência Taxonômica
ASIUM	Acquisition of Semantic knowledge Using Machine Learning Methods
BEM	Betweenness Measure
BNC	British Natural *Corpus*
CLA	Concept Lattices and Their Applications
CLUTO	Clustering Toolkit
CMM	Class Match Measure
CoNLL	Conference on Computational Natural Language Learning
CSC	Common Semantic Cotopy
DEM	Density Measure
DTD	Document Type Definition
ERCA	Eclipse's Relational Concept Analysis
FCA	Formal Concept Analysis
FCR	Família de Contextos Relacionais
FOIS	Formal Ontologies in Information Systems
FRR	Famíla de Reticulados Relacionais
FT	F-score Taxonômica
HTML	HyperText Markup Language
ICCS	International Conference on Conceptual Structures
ICFCA	International Conference on Formal Concept Analysis
ICS	Information Content Similarity
Illinois SRL	Illinois Semantic Role Labeler
JAIR	Journal of Artificial Intelligence Research
k-NN	k-Nearest Neighbor
LDC	Linguistic Data Consortium
LSDIS	Large Scale Distributed Information System
NCBI	National Center for Biotechnology Information
NLTK	Natural Language Toolkit
OWL	Web Ontology Language
PL	Precisão Léxica
POS	Part of Speech
PropBank	Proposition Bank
PT	Precisão Taxonômica
RLA	Relation Learning Accuracy
RCA	Relational Concept Analysis
SC	Semantic Cotopy
SOM	Self Organizing Map
SSM	Semantic Similarity Measure
SVM	Suport Vector Machine
TO	Taxonomic Overlap
W3C	World Wide Web Consortium
XMI	XML Metadata Interchange
XML	Extensible Markup Language

SUMÁRIO

1. INTRODUÇÃO

1.1 Motivação

Estruturas ontológicas como dicionários, tesauros, taxonomias e ontologias têm se tornado um importante recurso para sistemas de informação. Em sistemas de recuperação de informações, por exemplo, tais estruturas têm ajudado a minimizar problemas de vocabulário e recuperar informações mais relevantes. Essas estruturas provêem um suporte semântico que permite modificar a consulta do usuário, substituindo termos[1] desconhecidos do sistema por sinônimos, bem como enriquecendo esta consulta com termos relacionados pertencentes ao mesmo domínio.

Em tarefas de organização da informação, como classificação e agrupamento de documentos, essas estruturas conceituais[2] têm trazido ganhos de precisão. Elas têm sido usadas para melhorar tais tarefas enriquecendo os textos com novos conceitos do domínio [23]; auxiliando na determinação da classe ou grupo através da desambiguação dos termos dos documentos [12]; reduzindo a dimensionalidade do vetor de representação dos textos, através da substituição de conjuntos de termos pelos conceitos que os representam; e, ainda, ressaltando as características mais relevantes a partir das relações semânticas entre os termos [52], entre outras funções.

Embora seja grande a aplicabilidade das estruturas ontológicas, é alto o custo de sua construção e manutenção. Por esta razão, e devido ao grande volume e riqueza de documentos textuais digitais disponíveis atualmente, pesquisas têm sido realizadas com o objetivo de construir tais estruturas automaticamente a partir de textos. As abordagens propostas geralmente usam informações linguísticas e estatísticas sobre os termos no texto, para extrair os conceitos e as relações semânticas entre eles.

É comum, também, o uso de algoritmos de aprendizagem de máquina, principalmente de agrupamento, para identificar os conceitos a partir da coocorrência de termos em um contexto, e para construir as estruturas ontológicas propriamente ditas, usando abordagens hierárquicas [72].

Como a construção de estruturas ontológicas a partir de textos é considerada um problema de difícil solução, outras abordagens têm surgido. É o caso do método Formal Concept Analysis (FCA), apresentado no Capítulo 3, que foi introduzido na década de 80 para análise de dados [76] e que vem sendo aplicado na construção de tais estruturas [38, 171].

Um trabalho muito citado que segue essa abordagem é o de Cimiano, Hotho e Staab [41] que utilizam FCA para construir automaticamente ontologias a partir de textos. Cimiano *et al.* utilizam os argumentos sintáticos dos verbos (sujeito, objeto direto e objeto indireto) para caracterizar as entidades do domínio. E usam os próprios verbos para definir as relações entre essas entidades. Através do método FCA, que é fundamentado na teoria dos reticulados, as entidades são agrupadas conceitualmente e organizadas conforme os verbos com os quais se relacionam.

O método FCA tem sido usado para construir estruturas conceituais cujo propósito é apoiar diferentes tarefas em sistemas de informação. Os resultados relatados pelos pesquisadores são muito promissores. Ele tem sido aplicado em tarefas de classificação [56, 234], clusterização [23, 149], recuperação de informações [37, 44, 163], mineração de dados [211] e outras.

Como método para análise de dados ele também é interessante e trabalhos na área de linguística [155, 167] têm explorado esse aspecto. Nesse sentido, ele pode ser usado para apoiar a construção manual ou integrar uma das etapas em processos semiautomáticos de elaboração

[1] Entendemos "termo" como uma unidade atômica de significado. Pode ser representado por uma palavra, o radical de uma palavra, um sintagma ou n-grama.

[2] Neste documento, estruturas conceituais e estruturas ontológicas são tratadas como sinônimos.

de ontologias. Quanto às ontologias, mais especificamente, o método FCA tem sido usado para geração de axiomas [158], mapeamento, similaridade [45, 62] e aprendizagem de ontologias [17, 98, 69, 153, 182].

Embora sejam muitas as aplicações do método, os trabalhos têm explorado aspectos mais sintáticos ou estatísticos quando o objetivo é construir estruturas conceituais a partir de textos. Até encontramos algumas abordagens que consideram aspectos semânticos, mas poucos são os trabalhos que exploram, por exemplo, papéis semânticos de forma mais ampla. O trabalho de Valverde-Albacete [210] é um dos poucos, na atualidade, que utiliza FCA e papéis semânticos, no entanto, é voltado para análise linguística, tendo como propósito representar a FrameNet (Seção 4.4) através de reticulados de conceitos.

Os papéis semânticos (Capítulo 4), ainda que pouco explorados conjuntamente com o método FCA, têm sido utilizados com frequência em trabalhos na área de aprendizagem de ontologias a partir de textos [11, 14, 187, 216]. Nesse contexto, eles provêem informações semânticas relevantes na medida em que "expressam o significado dos argumentos dos verbos em situações descritas por esses verbos" [99]. Tais papéis permitem identificar, por exemplo, quem é o agente de uma ação - a entidade que provoca o evento caracterizado pelo verbo - e também quem é o paciente dessa ação - a entidade que sofre modificações em consequência do evento. Eles permitem, portanto, caracterizar melhor as relações entre as entidades do domínio, mesmo que tais entidades ocorram em diferentes posições sintáticas.

Além disso, os papéis semânticos imprimem aos seus argumentos restrições semânticas. Essas restrições definidas através de traços (concreto, abstrato, animado, ...) descrevem características das entidades que permitem identificar os argumentos que expressam conceitos do domínio e sugerir, de forma automática, uma organização mais adequada para esses conceitos.

Soma-se a isso o fato de verbos conseguirem representar relações não taxonômicas que são comuns em estruturas ontológicas de domínio. Assim como Navok e Hearst em [150], nós acreditamos que os verbos consigam capturar aspectos sutis de significado, sendo, portanto, importantes fontes de expressividade em tarefas de representação semântica.

Considerando que o método FCA permite construir estruturas conceituais de domínio cujos relacionamentos entre as entidades podem ser baseados nos verbos, e que os papéis semânticos permitem atribuir características aos conceitos extraídos dos textos, propomos uma abordagem de construção de estruturas ontológicas baseadas em FCA e em papéis semânticos.

Acreditamos que nossa abordagem sugira uma organização mais interessante do ponto de vista semântico, por enriquecer com papéis semânticos a intensionalidade dos conceitos formais. Além disso, por ser baseada em FCA, nossa abordagem pode ser gerada tanto de forma semiautomática como automática. Tal abordagem foi avaliada por medidas estruturais e funcionais. A medida estrutural de coesão lexical foi usada para analisar os conceitos gerados a partir da abordagem proposta. Também avaliamos nossos resultados na tarefa de categorização de textos[3], para a qual usamos medidas funcionais tradicionais.

A maioria dos estudos descritos nesse documento são voltados para a Língua Inglesa. Utilizamos *corpora* anotados semanticamente como Penn TreeBank Sample e SemLink 1.1 em nossa investigação. Utilizamos ainda os *corpora* WikiFinance e WikiTourism, respectivamente, dos domínios Finança e Turismo, para analisar nossa proposta em diferentes *corpora* e domínio. Esses *corpora* foram extraídos do WikiCorpus 1.0 e anotados com papéis semânticos pelo processador F-EXT-WS (Seção 5.3.4).

Realizamos ainda alguns estudos preliminares em Língua Portuguesa para a abordagem proposta, incluindo trabalhos em categorização de textos e extração de conceitos para o *corpus* PLN-BR CATEG.

Cabe mencionar que combinar estruturas FCA com papéis semânticos não é uma ideia

[3]Ao longo do documento, usaremos categorização e classificação como sinônimos; o mesmo vale para classificar e categorizar; e para classe e categoria.

recente. Kamphuis e Sarbo em [101], na década de 90, propõem a representação de uma frase em linguagem natural, associando tais elementos. Também na década de 90, o trabalho de Rudolf Wille [218] apresenta exemplos de estruturas FCA relacionadas a papéis semânticos. O objetivo de Wille, no entanto, é combinar grafos conceituais com estruturas FCA visando a formalização de uma lógica útil à representação e ao processamento de conhecimento.

Apesar de as abordagens daqueles autores parecerem promissoras à época em que foram propostas, até agora haviam sido pouco exploradas. Provavelmente devido à dificuldade de anotação dos textos, visto que o surgimento de etiquetadores automáticos de papéis semânticos é mais recente.

Mesmo com a profunda revisão bibliográfica realizada, não encontramos, até o momento, trabalhos que explorassem os papéis semânticos em conjunto com o método FCA para apoiar a construção de estruturas ontológicas a partir de textos. Dado que a proposta de combiná-los é pouco pesquisada, consideramos ser de interesse todo o esforço gerado ao investigá-la.

1.2 Objetivo

Nosso objetivo primário é combinar o método FCA com papéis semânticos para construir, de forma automática e a partir de informações textuais, estruturas ontológicas fortemente baseadas em relações não taxonômicas. A ideia é explorar as vantagens do FCA como método de agrupamento conceitual. O método FCA, quando comparado a outros métodos de agrupamento, permite delinear mais facilmente, do ponto de vista semântico, os grupos e subgrupos de uma hierarquia. Já no que tange aos papéis semânticos, investigamos o uso dessa informação semântica no processo de extração de informação e construção dos conceitos, sendo que nossa meta é utilizar tal informação para qualificar e, portanto, melhorar os grupos (conceitos) gerados pelo método FCA.

1.3 Desenvolvimento

Para alcançar o objetivo proposto, realizamos uma pesquisa de carácter fundamentalmente exploratório. Primeiramente, analisamos os *corpora* em Língua Inglesa, Penn TreeBank Sample e SemLink 1.1. Nesse estudo preliminar, procuramos identificar os sintagmas nominais, os papéis semânticos, as relações e as classes de verbos que poderiam contribuir para uma pesquisa mais expressiva do ponto de vista quantitativo (Seção 6.4). Os resultados dessa análise foram determinantes para o direcionamento de nossa pesquisa.

Em seguida, investigamos formas de incluir, nas estruturas FCA e suas extensões, as informações semânticas identificadas nos *corpora* analisados. Com esse fim, analisamos a adequação da estrutura Relational Concept Analysis (RCA), que é uma extensão do FCA, quanto à representação dos papéis semânticos em conceitos formais. Estudamos também as classes VerbNet 37.7 e 45.4, as quais foram indicadas, na pesquisa quantitativa, como as mais significativas para os *corpora* analisados (Capítulo 7). Neste estudo, investigamos a relação existente entre os papéis semânticos e as classes de verbos. Tal estudo nos permitiu visualizar e estabelecer a importância dessas classes para nossa pesquisa.

Ainda com o propósito de incluir informações semânticas em reticulados conceituais FCA, exploramos 6 casos de estudo. Nesses casos, propusemos formas de combinar verbos, papéis semânticos e classes de verbos em conceitos formais. Avaliamos os casos propostos, inicialmente, apenas sob um ponto de vista estrutural, a partir da medida de coesão lexical (Capítulo 8). Com base nos resultados provenientes dessa investigação, escolhemos os casos de estudo mais promissores para avaliar a aplicabilidade de nossa abordagem em outros *corpora* e domínios.

Para investigar essa aplicabilidade, realizamos, então, uma pesquisa aplicada na área de categorização de documentos, usando os *corpora* WikiFinance e WikiTourism. A meta era

analisar a efetiva contribuição das estruturas conceituais geradas a partir de nossa proposta na tarefa de classificação. Os resultados dessa investigação foram avaliados a partir de medidas funcionais usualmente aplicadas em tarefas de categorização (Capítulo 9).

Incluímos ainda em nossa pesquisa, estudos com um *corpus* para Língua Portuguesa, o *corpus* PLN-BR CATEG (Capítulo 9). Nesses estudos, abordamos as tarefas: extração de conceitos a partir de textos, categorização de documentos e, apesar da falta de recursos, a inclusão de papéis semânticos em estruturas FCA geradas a partir de textos.

1.4 Contribuição

Consideramos como uma de nossas principais contribuições o estudo que realizamos, quanto à inclusão de papéis semânticos nos contextos formais, a partir dos quais as estruturas FCA são geradas. O resultado desse estudo foi uma metodologia, aplicável em diferentes *corpora* e domínios, que permite criar estruturas FCA enriquecidas como papéis semânticos. Denominamos tal metodologia de Semantic FCA (SFCA) e a descrevemos, na forma de um procedimento, na Seção 10.2.

Embora a metodologia SFCA não gere, como resultado, ontologias tais como descritas na visão de Guarino [86], acreditamos em sua aplicabilidade. A abordagem é capaz de gerar estruturas conceituais a partir de textos de forma automática, estruturas estas que podem ser usadas para construir ontologias e apoiar tarefas na área de sistema de informação. As estruturas assim construídas são ontológicas e de domínio, e suas características cabem perfeitamente na definição de Gruber [85] para ontologias, definição esta que é aceita e utilizada no âmbito dessa investigação.

Julgamos igualmente relevante a pesquisa aplicada que realizamos na área de categorização de documentos. A partir dessa investigação, que gerou resultados muito animadores quanto à aplicabilidade das estruturas SFCA, pudemos também propor uma metodologia que descreve como usar essas estruturas na tarefa de classificação. Tal metodologia foi formalizada na Seção 10.3 deste documento.

Outra contribuição que consideramos importante refere-se à avaliação estrutural (coesão lexical), que propusemos e utilizamos para analisar as estruturas FCA construídas ao longo de nossa pesquisa. Consideramos tal avaliação relevante principalmente por existirem poucas medidas dessa natureza para analisar conceitos formais.

1.5 Organização

Este documento está organizado em 10 capítulos, seguidos das referências bibliográficas, assim constituídos:

- O Capítulo 2 introduz estruturas ontológicas, incluindo definições, classificações e elementos que geralmente as compõem. São mencionadas também as tarefas e as abordagens usadas em aprendizagem de ontologias a partir de textos, bem como métodos de avaliação e alguns ambientes existentes para construção semiautomática de ontologias.

- O Capítulo 3 descreve o método FCA com uma breve fundamentação matemática e aplicações. São mencionados, também, alguns dos algoritmos utilizados para gerar os reticulados de conceitos. Tecemos comentários, ainda, sobre medidas para avaliação de conceitos formais e sobre a extensão do método FCA: Relational Conceptual Analysis.

- O Capítulo 4 aborda os papéis e etiquetadores semânticos bem como recursos lexicais utilizados nas tarefas de anotação.

- O Capítulo 5 descreve de forma sucinta os *corpora*, as ferramentas e as bases lexicais e ontológicas utilizadas nos estudos exploratórios.

- O Capítulo 6 descreve nossa questão e método de pesquisa. Apresentamos ainda, nesse capítulo, o pré-processamento dos *corpora* Penn TreeBank Sample e SemLink 1.1, bem como as heurísticas usadas para extração de sintagmas nominais e informações preliminares extraídas desses *corpora*.

- O Capítulo 7 apresenta o primeiro estudo que realizamos no âmbito de nossa pesquisa, no qual analisamos a extensão RCA e algumas classes de verbos.

- O Capítulo 8 descreve nossa investigação quanto à forma de incluir as informações semânticas extraídas dos textos nos contextos formais usados para gerar estruturas FCA.

- O Capítulo 9 descreve o estudo realizado na área de categorização de textos, que visa avaliar a aplicação de nossa proposta em outros *corpora* e domínios. Para esse fim, usamos textos da Wikipédia dos domínios de Finanças e Turismo para Língua Inglesa. É neste capítulo que relatamos algumas das pesquisas realizadas com textos em Língua Portuguesa.

- E o Capítulo 10 apresenta as considerações sobre o estudo realizado e as metodologias resultantes de nossa investigação exploratória: a metodologia para construção de estruturas SFCA e a metodologia para categorização de documentos baseada em SFCA. Abordamos também nesse capítulo, os trabalhos futuros e as publicações realizadas durante o doutoramento.

2. ESTRUTURAS ONTOLÓGICAS E APRENDIZAGEM A PARTIR DE TEXTO

Este capítulo faz uma introdução a estruturas ontológicas e à aprendizagem dessas estruturas a partir de texto. São apresentados conceitos e classificações de estruturas ontológicas, nos quais se destaca o uso do termo "ontologia" para nomear estruturas conceituais com diferentes graus de expressividade semântica. São comentadas também as tarefas relacionadas à aprendizagem de ontologias, as abordagens mais comumente usadas para extrair conceitos e seus relacionamentos, alguns ambientes de apoio à aprendizagem a partir de textos e também formas de avaliação de ontologias.

2.1 Definições de ontologia

Em filosofia, a palavra de origem grega "Ontologia" é usada para designar um dos ramos da Metafísica que estuda a essência do "ser" ou da existência [38, 84]. Nesse contexto, Ontologia se preocupa com as características das entidades reais, provendo-lhes conceitos (definições) e identificando aspectos essenciais quanto ao tipo e à estrutura dos objetos, das propriedades, dos eventos, dos processos e das suas relações exatamente como elas existem no mundo [197]. As ontologias descrevem os tipos de fenômenos que podem existir, tais como pessoas, lugares, eventos, relacionamentos, etc [217]. Têm como objetivo, ainda, prover a catalogação das entidades, definindo classes, hierarquias e relações classes-instâncias em conformidade com a realidade [53].

Ainda no sentido filosófico, Guarino em [86] diferencia "Ontologia" (com "O" maiúsculo) de "ontologia" (com "o" minúsculo). No primeiro caso, refere-se à disciplina da Metafísica e no segundo, a um sistema de categorias que expressa uma visão de mundo, tal como a "ontologia de Aristóteles".

Já em ciência da computação, a palavra "ontologia" tem sido usada para expressar um modelo conceitual, recebendo, segundo Guizzardi *et al.* [88], duas interpretações: uma conforme a comunidade de modelagem conceitual e outra de acordo com as comunidades de inteligência artificial, engenharia de software e web semântica.

Em modelagem conceitual, ontologia aproxima-se do sentido filosófico, visto que procura definir "um sistema de categorias formais independente de domínio", mas fundamentado na realidade e que pode ser usado para modelar diferentes domínios [88]. Guarino em [86] também define ontologias como modelos conceituais. Para Guarino, uma ontologia aproxima, no sentido intencional, o modelo conceitual pretendido para um determinado domínio. É em [86] que Guarino introduz o termo "ontologia formal", conceituando ontologia como "uma teoria lógica que expressa o significado atribuído a um vocabulário formal".

A interpretação de Guarino para ontologias é ilustrada na Figura 2.1. Através de uma linguagem L, pode-se gerar diferentes modelos $M(L)$ para representar a conceituação de um domínio. Uma conceituação é uma abstração, uma "visão de mundo" sobre um domínio. Uma determinada conceituação C é expressa segundo um "compromisso ontológico" K. Um "compromisso ontológico" estabelece uma terminologia (vocábulos em L) cuja semântica é definida (ou restringida) por axiomas.

É o "compromisso ontológico" que determina a intenção do modelo, ou seja, especifica o significado dos termos e de suas relações, caracterizando, assim, a conceituação pretendida para aquela representação. A ontologia é uma aproximação do modelo pretendido (identificado na Figura 2.1 como I_k), pois podem existir entidades no domínio que a teoria lógica não é capaz de representar de forma fiel à realidade [86].

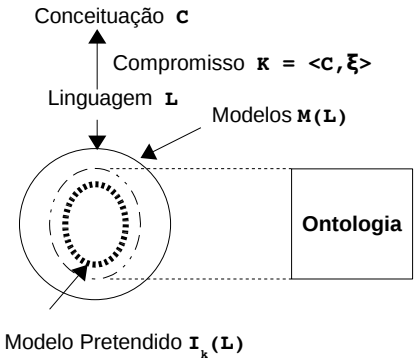


Figura 2.1 – Ontologias como aproximação de modelos (adaptado de [86])

Já nas demais comunidades mencionadas - inteligência artificial, engenharia de software e web semântica - ontologia é considerada um "artefato concreto de engenharia". Ela é uma estrutura formal composta por conceitos e relações entre conceitos. Inclui também um conjunto de axiomas que restringem a interpretação dessa estrutura, bem como refletem o conhecimento sobre um domínio no contexto de uma aplicação. O conhecimento representado refere-se a "casos" (situações) relevantes para a aplicação [87]. Isso significa que o comprometimento da ontologia com a realidade (fundamentação) nem sempre é o objetivo principal, mas quando aparece é definido em função apenas das necessidades da aplicação [88, 86].

Para essas comunidades, conforme Gómez-Pérez em [160], a definição mais citada é a de Gruber, na qual ontologia é "uma especificação explícita de uma conceituação" [85]. Segundo Gruber, conceituação nada mais é do que uma abstração (visão simplificada) do "mundo" relevante para uma determinada aplicação. Nesse contexto, a ontologia tem como objetivo compartilhar conhecimento, expressando as entidades dessa abstração de "mundo" segundo a interpretação consensual de um grupo de pessoas e não apenas de um indivíduo [53].

2.2 Categorias de estruturas ontológicas

Apesar de a definição de Gruber ser a mais aceita entre os cientistas da computação, ela tem sido questionada devido a sua abrangência [84, 87]. Conforme Smith e Welty em [198], essa definição permite interpretações nas quais ontologias podem ser basicamente qualquer estrutura conceitual, desde simples catálogos até teorias axiomatizadas. Embora existam muitas críticas quanto ao uso do termo ontologia para denominar estruturas conceituais como taxonomias, por exemplo, os pesquisadores têm procurado distinguir as "ontologias" quanto ao que alguns chamam de expressividade semântica [26, 55], outros de precisão ou, ainda, de complexidade no tange à capacidade de raciocínio [198, 222].

A Figura 2.2 apresenta a classificação de ontologias segundo Smith e Welty em [198], que é uma das primeiras iniciativas neste sentido. Essa classificação, que não é exaustiva, mas acreditamos abranger boa parte das estruturas conceituais, inclui: catálogos que definem uma lista finita de termos de um domínio (vocabulário controlado); arquivos-texto de um mesmo domínio; glossários nos quais o significado dos termos é descrito em língua natural (como dicionários); tesauros contendo definições de termos e seus relacionamentos; hierarquias que organizam os termos em taxonomias usando a relação *is-a* (é-um) ou em estruturas utilizando outras relações de ordem como em FCA; *frames* que organizam as entidades do domínio a partir de suas classes e propriedades; e estruturas conceituais que, além de descrever conceitos e seus

relacionamentos, incluem axiomas que restringem a semântica desses conceitos, permitindo a realização de inferências (raciocínio) [26].

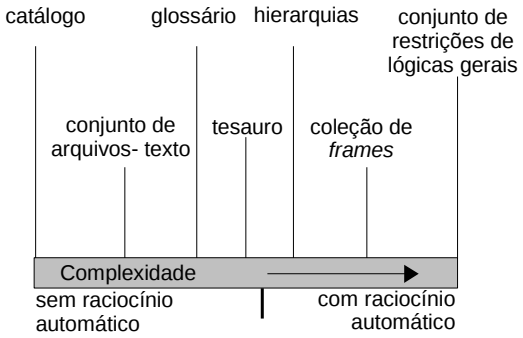


Figura 2.2 – Classificação de ontologias quanto à complexidade (adaptado de [198]).

De acordo com Guarino em [86], todas as estruturas da Figura 2.2 podem ser ditas ontológicas visto que a definição de conceituação de Gruber, em [85], permite interpretações de caráter puramente extensional. Gruber define conceituação como uma estrutura $< D, R >$, onde D é o domínio e R é o conjunto de relações relevantes em D, ou seja, R considera relações que refletem casos (estados particulares de mundo) [85].

Guarino atribui essa flexibilidade na definição à ausência de uma noção de conceituação mais formal, focada no significado, que determine, além das relações extensionais, relações intensionais [86]. Esse é um fator que tem sido apontado como um dos entraves aos avanços na área de interoperabilidade de sistemas pois, sem uma semântica formalmente descrita e clara (bem definida), o casamento (*matching*) de ontologias restringe-se ao casamento de vocábulos (*strings*) e seus relacionamentos. Relacionamentos estes que nem sempre descrevem "visões" equivalentes de mundo, embora pertençam a um mesmo domínio [176].

Guarino em [86] também propõe uma classificação para as estruturas ontológicas, no entanto considerando aspectos mais gerais (Figura 2.3). Ele agrupa as estruturas em:

- ontologias de alto nível - são aquelas cujo propósito é descrever conceitos genéricos independentes de domínio e que podem ser reusadas na construção de novas ontologias; descrevem conceitos como espaço, tempo, eventos, etc.

- ontologias de domínio - utilizam um vocabulário pertencente a um determinado domínio, especializando conceitos genéricos definidos em ontologias de alto nível.

- ontologias de tarefa - fazem uso de um vocabulário relacionado à execução de tarefas ou atividades genéricas e também especializam conceitos de ontologias de alto nível.

- ontologias de aplicação - descrevem o vocabulário de uma aplicação específica, na qual os conceitos geralmente representam os papéis que as entidades de um domínio desempenham durante a execução de uma tarefa ou atividade.

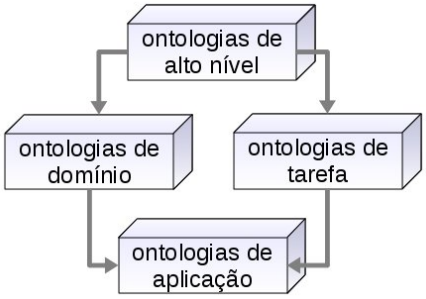


Figura 2.3 – Classificação de ontologias segundo Guarino (adaptado de [86])

2.3 Primitivas de representação de ontologias

De acordo com Gruber em [84], uma estrutura ontológica, no contexto da ciência da computação e da informação, estabelece "um conjunto de primitivas de representação com o qual é possível modelar um domínio de conhecimento ou de discurso. Essas primitivas são tipicamente classes (ou conjuntos), atributos (ou propriedades) e relacionamentos (ou relações entre os elementos das classes)". Segundo o autor ainda, as primitivas descrevem informações relativas ao significado das entidades do domínio, incluindo restrições lógicas a partir das quais mecanismos de raciocínio podem atuar de forma consistente em uma aplicação [84].

Gómez-Pérez em [160], baseando-se no trabalho de Gruber [85], detalha tais primitivas de representação, organizando-as em conceitos (que são as classes mencionadas por Gruber), instâncias, relações, funções e axiomas. Seguimos a organização de Gómez-Pérez, ao detalhar essas primitivas, à exceção de funções, as quais descrevemos, a exemplo de outros autores, como um tipo de relação. Desta forma, o conhecimento acerca de um domínio pode ser representado em uma estrutura ontológica por meio de:

- **Conceitos** (classes) - são as principais entidades de uma ontologia e representam um conjunto de indivíduos do domínio. "Podem ser concretos, elementares ou compostos, reais ou fictícios. Podem também descrever tarefas, funções, ações, estratégias, processos de raciocínio, etc" [160] . *Livro* e *Autor* são exemplos de conceitos.

- **Instâncias** (indivíduos ou objetos) - são elementos particulares associados a conceitos do domínio [160]. Por exemplo, *Os Maias* representa um livro específico e *Eça de Queirós*, um autor. Eles são instâncias, respectivamente, das classes *Livro* e *Autor.*

- **Relações** - estabelecem ligações n-árias entre os conceitos do domínio [160]. São classificadas geralmente em relações taxonômicas ou não taxonômicas. As relações taxonômicas são binárias e permitem a organização dos conceitos em uma hierarquia. Desta forma, usando como ordem a inclusão, conceitos mais abrangentes são relacionados a conceitos cada vez mais específicos, formando uma taxonomia. As relações taxonômicas são consideradas por Jurafsky e Martin [99] como um subtipo de hiperonímia[1]. As relações é-um (*is-a*) ou de hiponímia[2] também são usadas em taxonomias. Já as relações não taxonômicas, também conhecidas como transversais, podem relacionar os conceitos sob diferentes aspectos, e não formam necessariamente hierarquias e nem sempre são biná-

[1]A relação lexical de hiperonímia, também chamada de "é-superior-a", "é-superclasse-de" (*superordinate*), denota que um conceito é superclasse de outro [99]. Portanto, nessa relação binária, existe um item que funciona como um "protótipo" (uma classe) que representa a generalização de outros itens mais específicos. Por exemplo, "ave" é hiperônimo de "canário" e "pato" [229].

[2]A hiponímia é inversa à hiperonímia. Nesse caso, "canário" é hipônimo de "ave" [229].

rias. São exemplos de relações não taxonômicas as relações de sinonímia[3], de agregação ou composição como *é-parte-de* (meronímia[4]/holonímia[5]), associação como *conectado-a*, etc. Gómez-Pérez observa que algumas relações não taxonômicas são, na verdade, relações funcionais, com a relação *é-mãe-de*. Segundo Wong *et al.* [222] e Priss [168] esse é o caso também dos papéis semânticos (Capítulo 4), os quais expressam as funções atribuídas pelos verbos aos conceitos presentes em uma sentença. É importante mencionar que, para alguns autores, a relação de meronímia é considerada taxonômica, pois permite descrever um conceito a partir de suas partes, viabilizando também a construção de uma hierarquia [168]. Cabe ressaltar ainda que algumas relações definidas por Gómez-Pérez são denominadas **atributos** ou **propriedades** por outros autores [38], pois relacionam conceitos a tipos de dados (inteiro, real, *string*,...), tais como *preço* e *número-de-páginas* de um livro.

- **Axiomas** - são sentenças lógicas que definem a semântica dos conceitos e das suas relações. Atuam de forma restritiva, excluindo interpretações [85].

Obviamente que, no contexto da ciência da computação, as primitivas precisam ser "interpretáveis" por máquina e por esta razão são usadas, na representação de ontologias, linguagens descritivas baseadas geralmente em lógica de primeira ordem [84]. De acordo com Euzenat e Schvaiko em [55], ontologias normalmente são representadas em OWL (Web Ontology Language), sendo esta linguagem recomendada, inclusive, pela World Wide Web Consortium(W3C[6]).

Diferentemente de outras áreas, no que se refere à interpretação de ontologias, a ciência da computação costuma ser mais pragmática quanto à limitação das estruturas de representação usadas para caracterizar os elementos de estruturas ontológicas. Como nem tudo pode ser representado pelas linguagens usadas para descrever ontologias, "o que existe no mundo da aplicação é o que pode ser representado e manipulado computacionalmente". Esse costuma ser o ponto de vista dos pesquisadores de inteligência artificial [85].

Mesmo com a limitação imposta pelas estruturas de representação computáveis, construir ontologias com qualidade ainda exige muito esforço manual. Uma das formas de reduzir esse esforço é construir sistemas computacionais capazes de extrair de fontes de dados, como textos, as primitivas descritas por Gómez-Pérez. A área que se preocupa como o desenvolvimento de tais sistemas é chamada de aprendizagem de ontologias, e é esta área o assunto da próxima seção.

2.4 Abordagens para aprendizagem de ontologias

De acordo com Cimiano em [38], o termo 'aprendizagem de ontologia' (*ontology learning*) foi originalmente usado por Maedche e Staab em [134] para descrever o processo de "aquisição de um modelo de domínio a partir de dados". Esse processo, que tem natureza multidisciplinar, auxilia a engenharia semiautomática e cooperativa de ontologias. Nele, diferentes disciplinas atuam de forma complementar, trabalhando com diferentes tipos de dados, sendo estes não estruturados, semiestruturados ou totalmente estruturados [134].

Já Gómez-Pérez e Manzano-Macho em [80] definem o termo 'aprendizagem de ontologia' como a aplicação de um conjunto de métodos e técnicas para construção de ontologias, que

[3]A relação lexical de sinonímia entre dois conceitos estabelece que eles são substituíveis um pelo outro em qualquer sentença sem modificar o sentido dessa sentença [99].

[4]A meronímia expressa uma relação de inclusão semântica entre duas unidades lexicais, uma denotando a parte (merônimo) e outra referenciando a um todo (holônimo). Por exemplo, "dedo" é merônimo de "mão" [229].

[5]Holonímia é a relação inversa à meronímia. Por exemplo, "mão" é holônimo de "dedo" [229].

[6]http://www.w3.org/TR/owl-features/

permitem enriquecer ou adaptar de forma semiautomática ontologias existentes usando, para isso, fontes de informação e conhecimento heterogêneas. Tal processo, segundo os autores, reduz o tempo e esforço necessário para o desenvolvimento de ontologias.

Maedche e Staab em [134] classificam o processo conforme a natureza dos dados, pois a estrutura dos dados geralmente acaba por definir os métodos e técnicas adequados para a extração do modelo de domínio. Eles organizam as abordagens em:

- aprendizagem de ontologia a partir de textos - combina métodos linguísticos, estatísticos, baseados em padrões e de aprendizagem de máquina para extrair o modelo do domínio a partir de textos não estruturados;
- aprendizagem de ontologia a partir de dicionários - faz uso de métodos linguísticos, baseados em padrões e estatísticos, para definir o modelo a partir de dicionários interpretáveis por máquina;
- aprendizagem de ontologia a partir de bases de conhecimento - aplica métodos indutivos e estatísticos em bases de conhecimento, obtendo o modelo a partir das relações existentes entre os elementos descritos nessas bases;
- aprendizagem de ontologia a partir de esquemas semiestruturados - usa métodos de aprendizagem de máquina para definir ontologias a partir de recursos que possuem uma estrutura definida, como Extensible Markup Language (XML) ou Document Type Definition (DTD); e
- aprendizagem de ontologia a partir de esquemas relacionais - explora a correlação dos dados usando técnicas de engenharia reversa para extrair ontologias de bancos de dados [79].

Como o foco deste trabalho é o primeiro caso - aprendizagem de ontologia a partir de textos - apenas esta abordagem será detalhada.

2.5 Aprendizagem de ontologia a partir de textos

Aprendizagem de ontologia a partir de texto é o processo de identificar termos, conceitos, relações e opcionalmente axiomas a partir de informações textuais com o objetivo de usar tais elementos na construção e manutenção de ontologias [222].

A primeira etapa desse processo de aprendizagem é a escolha do *corpus* que será usado como fonte de dados para construção da estrutura conceitual. A qualidade e a riqueza do *corpus* são fundamentais para o bom desempenho de qualquer abordagem de extração de informações a partir de textos. O *corpus* pode ser composto por textos não estruturados (em formato livre), com algum formato (HTML, XML, ...) ou, ainda, com algum tipo de anotação morfossintática ou semântica.

Como as ontologias modelam domínios, é comum que os *corpora* sejam definidos especialmente para construí-las. Por esta razão, a escolha do *corpus*, tradicionalmente, cabe ao usuário [105, 106] ou ao especialista do domínio [7].

Em abordagens automáticas ou semiautomáticas, o mais usual é constituir o *corpus* a partir de documentos *web* [100, 131, 187, 216, 222] ou a partir de recursos baseados em *web* como a Wikipédia [46, 113]. A *web* é considerada um importante recurso em tarefas de aquisição de conhecimento pois, além de ser um grande repositório de informações, que atende a inúmeros domínios, ela permite, em razão da redundância, medir a relevância e a confiabilidade dessas informações [186]. Já o uso da Wikipédia se popularizou por prover as informações de forma semiestruturada, possuir um vocabulário rico, ter capacidade de atualização e, principalmente,

por ser um recurso cuja natureza é mais próxima à das ontologias do que à dos textos livres [203].

A construção de *corpora* a partir da *web* é realizada com o auxílio de motores de busca. Para isso, são submetidos aos motores de busca um glossário de termos previamente definido ou apenas alguns termos (ou palavras-chave), ditos "sementes" (*seed-words*), que caracterizam os conceitos mais gerais da estrutura conceitual a ser construída [187]. Nesse último caso, geralmente, o *corpus* é formado paulatinamente. Cabe ao especialista disparar o processo, informando as sementes iniciais. A partir dessas sementes é formado um *corpus* com documentos *web*. Desse *corpus*, são extraídos conceitos relacionados às sementes iniciais, os quais passam a ser as novas sementes, reiniciando o processo [222].

Já quando a aplicação é o enriquecimento de ontologia, a seleção é feita a partir de conceitos dessa ontologia [1]. Neste caso, as consultas feitas aos motores de busca procuram capturar a semântica, associando sinônimos, hiperônimos, atributos e outros elementos relacionados a esses conceitos na ontologia [1].

Cabe ressaltar, ainda, que muitos experimentos de aprendizagem de ontologias a partir de textos são realizados com *corpora* jornalísticos, no entanto Jia *et al.* [98] observam que textos científicos podem ser mais indicados pois descrevem os conceitos em diferentes níveis de abstração, sendo, assim, naturalmente hierárquicos.

Constituído o *corpus*, a próxima etapa é a construção da estrutura ontológica, iniciando, desta forma, o processo de aprendizagem propriamente dito. Para extrair dos textos as primitivas de representação e estruturá-las adequadamente, é necessária a realização de várias tarefas. Essas tarefas e as abordagens usadas para realizá-las são os assuntos das próximas subseções. Na sequência, são comentados também alguns trabalhos que descrevem ambientes de aprendizagem de ontologias a partir de textos, e, ainda, as dificuldades inerentes ao processo de aprendizagem.

2.5.1 Tarefas em aprendizagem de ontologia a partir de textos

O processo de aprendizagem de ontologia a partir de textos envolve diferentes tarefas as quais Cimiano em [38] descreve como um "bolo em camadas" (Figura 2.4). Essas camadas, que detalhamos a seguir, iniciam com a identificação dos termos (camada base) e vão até a geração de axiomas gerais (camada topo).

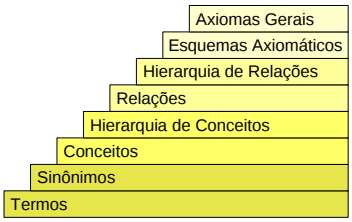


Figura 2.4 – Tarefas em aprendizagem de ontologia (adaptado de [38])

- **Termos** - a extração de termos é a primeira tarefa do processo de construção de uma ontologia e consiste em encontrar os termos ou representações simbólicas relevantes para conceitos e relações.
- **Sinônimos** - a tarefa de identificação de sinônimos consiste em encontrar palavras que denotem os mesmos conceitos, ou seja, que compartilhem um significado.
- **Conceitos** - a formação de conceitos deve incluir preferencialmente uma definição intensional, prover extensões dos conceitos e, ainda, representações simbólicas usadas para referenciá-los [29].

- **Hierarquia de conceitos** - a construção da hierarquia inclui indução, extensão e refinamento da estrutura conceitual. A tarefa de indução corresponde à geração da hierarquia propriamente dita, na qual os conceitos são organizados segundo uma relação de ordem, formando um semireticulado superior[7]. A tarefa de refinamento consiste em estender a hierarquia incluindo subconceitos. E a tarefa de extensão se destina a encontrar novas representações lexicais para um mesmo conceito.

- **Relações e Hierarquia de Relações** - a identificação de relações, muitas vezes, restringe-se a relações binárias e inclui tarefas como: encontrar os conceitos que possuem algum tipo de relação não taxonômica; especificar essas relações, propondo rótulos e definições adequadas; e aprender uma ordem hierárquica para essas relações.

- **Esquemas Axiomáticos** - essa tarefa faz uso de um sistema de axiomas já existente que define, por exemplo, axiomas de equivalência e disjunção para conceitos, e também axiomas que descrevem as propriedades das relações, como transitividade, simetria, etc. A tarefa consiste em instanciar axiomas, ou seja, identificar quais conceitos e relações estabelecem tais axiomas.

- **Axiomas gerais -** essa tarefa consiste em aprender axiomas que estabelecem relacionamentos mais complexos entre conceitos e relações.

Para suportar essas tarefas, uma variedade de recursos e técnicas tem sido usada. As abordagens usadas para esse fim são o assunto da próxima subseção.

2.5.2 Abordagens para aprendizagem de ontologias a partir de textos

A aplicabilidade das ontologias na solução de diferentes problemas, o volume de textos disponíveis, principalmente na *web*, e o grande esforço manual ainda necessário para construção de tais estruturas, têm aumentado o interesse na área de aprendizado de ontologias a partir de textos. Tanto que, na última década, diferentes abordagens com esse fim foram propostas. Nessas abordagens, é comum a combinação de técnicas linguísticas, estatísticas e de aprendizagem de máquina com recursos lexicais.

Nas próximas subseções são apresentadas as abordagens para aprendizagem de ontologia conforme a organização proposta por Gómez-Pérez e Manzano-Macho em [80], que as classificam em abordagens baseadas em técnicas linguísticas, estatísticas e de aprendizagem de máquina. Cabe ressaltar que descrevemos apenas as abordagens mais usuais na bibliografia pesquisada.

2.5.2.1 Técnicas linguísticas

As técnicas linguísticas são dependentes do idioma das sentenças do *corpus*, pois se baseiam em características morfológicas, sintáticas ou semânticas identificadas nos textos [80]. Nas abordagens estudadas, é muito comum o uso dessas técnicas no pré-processamento dos documentos. Elas são utilizadas para:

- segmentação do texto (*tokenization*) - os sinais de pontuação ajudam a identificar os limites de uma sentença. É nessa fase que as sentenças são segmentadas em termos [38].

[7] Quando os conjuntos ordenados não possuem supremo ou infímo, são chamados de semirreticulados. Se a estrutura possuir apenas o supremo denomina-se semirreticulado superior; se possuir apenas o ínfimo, semirreticulado inferior [48]. Mais informações sobre conjuntos ordenados e reticulados podem ser encontradas no Apêndice A.

- normalização dos termos - pode incluir a identificação e transformação de datas e indicadores de tempo, por exemplo, em formatos padronizados; a expansão de abreviaturas presentes nos textos; e a normalização morfológica através de processos de conflação como *stemming*[8] e lematização[9]. A redução das variações linguísticas dos termos é importante, principalmente quando abordagens linguísticas e estatísticas são combinadas para determinar os termos mais relevantes [93].

- etiquetagem com rótulos de *Part-Of-Speech* (POS) - consiste em etiquetar os termos com as suas classes gramaticais (substantivo, verbo, adjetivo, advérbio, ...). As classes gramaticais ajudam a identificar o tipo das entidades do domínio. Por exemplo, substantivos podem ser classes, adjetivos podem indicar atributos e verbos podem descrever eventos, estados, ações, etc [38, 93].

- reconhecimento de entidades nomeadas - consiste em encontrar no texto os termos que identificam os objetos no mundo, incluindo nomes de pessoas, organizações, lugares, etc. As entidades nomeadas tipicamente correspondem às instâncias em uma ontologia [38].

- análise sintática - consiste em identificar, nas sentenças, elementos sintáticos como sujeito, objeto direto, etc. Com esse propósito podem ser aplicados *parsers*, que geram a árvore sintática completa das sentenças, ou *chunk parsers*, que realizam análises parciais abrangendo apenas alguns elementos sintáticos considerados mais importantes conforme a tarefa em questão [38, 93]. É usada para apoiar diferentes tarefas de aprendizagem de ontologias, principalmente no que se refere à identificação de relações entre os termos, tanto de ordem taxonômica [91] quanto não taxonômica [38].

- identificação de contextos semânticos - é comum o uso da hipótese de distribuição semântica de Harris, que relaciona o significado de um termo com o contexto no qual ele aparece. Esse contexto pode ser "gráfico", delimitado por uma "janela", que pode ser, por exemplo, uma sentença, um parágrafo, um número determinado de palavras imediatamente anteriores ou posteriores ao termo, ou mesmo o documento inteiro. O contexto pode ser também "morfossintático", como nos trabalhos de Hindle [92] e Grefenstette [83]. Nesse último caso, alguns autores denominam o contexto de *topic signatures* [1] [3]. Combinados com abordagens estatísticas, os contextos podem ser usados para identificar relações de sinonímia. Já em algoritmos de agrupamento, são usados para definir os conceitos de uma estrutura ontológica. Um conceito corresponde a um grupo de termos de um mesmo contexto semântico.

- identificação de relações semânticas: As relações podem ser identificadas também a partir de padrões lexicossintáticos, como os de Hearst [91] (Figura[10] 2.5), por exemplo, que permitem extrair relações de hiperonímia e hiponímia do texto. De acordo com Cimiano em [38] esses padrões são mais eficientes em grandes *corpora,* visto que em *corpora* menores são mais difíceis de ocorrer. A WordNet também é muito usada para identificar relações, visto que ela provê, além de relações de sinonímia e antonímia[11], relações de hiperonímia, hiponímia, meronímia e holonímia. Outras relações semânticas de natureza transversal têm sido detectadas e rotuladas a partir de verbos [192, 216, 222] e, mais

[8] *Stemming* elimina as terminações das palavras, reduzindo-as a uma cadeia que, em alguns casos, corresponde ao seu radical. Palavras como "belo", "bela" e "beleza", por exemplo, seriam representadas por "bel" [99].

[9] Lematização converte os termos para sua forma canônica, ou seja, os verbos vão para o infinitivo e os substantivos vão para masculino-singular, se existir. No caso da lematização, as palavras "belo", "bela" e "beleza" seriam representadas pelo lema "belo" [99].

[10] A etiqueta NP identifica sintagmas nominais.

[11] A relação lexical de antonímia é inversa à sinonímia. Dois conceitos são antônimos se possuem sentidos opostos [99].

recentemente, a partir de papéis semânticos [11]. No caso desses últimos, é necessário o uso de etiquetadores de papéis semânticos. Esses etiquetadores são abordados na Seção 4.7 deste documento.

- desambiguação de sentido - procura resolver a ambiguidade de significado de palavras ou frases. Geralmente utiliza recursos lexicais como a WordNet para apoiar essa tarefa [1]. *Topic signatures,* associadas com medidas para calcular a distância semântica como Dice, Jaccard, cosseno, informação mútua e outras, também são usadas [38].

- resolução de correferência: consiste em identificar as várias formas de referência a um mesmo objeto no texto. Nesses casos, analisadores sintáticos (*parsers*) completos podem ser necessários.

$$NP \text{ such as } \{NP,\}^*\{(or \mid and)\}\ NP$$

$$such\ NP \text{ as } \{NP,\}^*\{(or \mid and)\}\ NP$$

$$NP \text{ including } \{NP,\}^*NP\ \{(or \mid and)\}\ NP$$

$$NP \text{ especially } \{NP,\}^*\{(or \mid and)\}\ NP$$

Figura 2.5 – Alguns padrões definidos por Hearst [91].

2.5.2.2 Técnicas estatísticas

As técnicas estatísticas são aplicadas geralmente em conjunto com as técnicas linguísticas e com algoritmos de aprendizagem de máquina. Têm o papel comumente associado a critérios de relevância, sendo usadas frequentemente para identificar os conceitos [1] e as relações mais expressivas entre conceitos no *corpus* [80].

Uma das medidas mais citadas na literatura é a tf-idf [185]. Ela é usada para identificar conceitos em vários trabalhos [13, 43, 64, 230]. O método C-Value/NC-Value [66] também tem sido muito referenciado [43, 180]. Segundo Spasic *et al.* [200], ele é adequado para extrair multitermos, pois combina conhecimento linguístico e estatístico ao definir a importância dos termos. Há ainda trabalhos com essa finalidade que utilizam qui-quadrado [30] e informação mútua [41, 121, 187], dentre outras medidas.

Por outro lado, há pesquisadores como Navigli *et al.* que propõem suas próprias medidas de relevância. Os autores em [151] combinam duas medidas para identificar os termos mais significativos para um domínio: escore de relevância e consenso. O escore de relevância mede a quantidade de informação capturada a partir de um *corpus* de domínio em relação aos *corpora* usados no processo de aprendizagem. A medida de consenso ajuda a escolher, dentre os termos identificados, aqueles que aparecem com mais frequência nos documentos do domínio.

Em outros trabalhos, como o de Yang e Callan em [225], a máquina de busca do Google tem sido usada para identificar os conceitos multitermos mais relevantes. São considerados importantes aqueles termos que ocorrem com maior frequência (acima de um determinado limiar).

Uma abordagem que vem ganhando espaço é a híbrida. Zhang *et al.* em [233] e Butters e Ciravegna em [31] apresentam resultados animadores ao combinarem mais de uma medida de ponderação dos termos. Butters e Ciravegna em [31] propõem também o uso de limiares dinâmicos, baseados na média e no desvio padrão das medidas. De acordo com os autores, em

geral, limiares fixos não levam em consideração a real distribuição dos termos. Vale mencionar que, em uma de nossas publicações (Seção 9.2.3), usamos tal abordagem, combinando as medidas tf-idf e C-Value para extrair conceitos de textos em Língua Portuguesa [145].

Cimiano *et al.* em [41] extrai relações não taxonômicas entre verbos e seus argumentos explorando medidas como probabilidade condicional e informação mútua. Lame e Desprès [118] também utilizam informação mútua, mas para identificar relações entre conceitos a partir de seus contextos.

Já no caso dos algoritmos de aprendizagem de máquina, as medidas são mais comuns em agrupamentos sendo usadas para estabelecer a similaridade entre conceitos a partir de seus contextos [135].

2.5.2.3 Técnicas de aprendizagem de máquina

As técnicas de aprendizagem de máquina associadas com medidas de distância semântica, como Dice, Jaccard, cosseno, informação mútua e outras, são usadas em diferentes tarefas do processo de aprendizagem de estruturas ontológicas. Elas são utilizadas principalmente para:

- Aprendizagem de conceitos: Vários algoritmos com diferentes paradigmas de aprendizagem são usados para esse propósito. An e Chen em [5], por exemplo, propõem melhoramentos no algoritmo Naïve Bayes com o objetivo de descobrir regras no *corpus* que possam ser usadas para identificar conceitos. Chin *et al.* [34] também usam o Naïve Bayes para encontrar conceitos, no entanto seu método consiste em agrupar as palavras em categorias pré-definidas conforme as suas probabilidades de distribuição. Fortuna *et al.* em [65] usam o algoritmo k-Means para descobrir tópicos que possam caracterizar conceitos em uma estrutura ontológica. Já em [64], esses autores utilizam aprendizagem ativa com base em classificadores SVM (Support Vector Machine) para identificar novos conceitos.

- Extração de relações e geração de hierarquias conceituais: é mais frequente o uso de algoritmos com abordagem não supervisionada. Dittenbach *et al.* [50], por exemplo, utilizam redes neurais SOM (Self-Organizing Map) para agrupar termos relevantes e descobrir relacionamentos como o de sinonímia. Maedche e Staab em [132] identificam relações conceituais binárias usando regras de associação. Karoiu *et al.* [103] constroem a estrutura ontológica a partir do algoritmo k-Means. E, Bisson *et al.* [21], com o mesmo propósito usam algoritmos de agrupamento conceitual. É importante comentar que, nos últimos anos, o método FCA (Capítulo 3) também passou a ser usado como método de agrupamento conceitual em aprendizagem de ontologias [16, 42, 72, 89].

- Enriquecimento de ontologias: a classificação de conceitos é frequentemente relacionada à tarefa de popular ontologias. Faz uso em geral de algoritmos com abordagem supervisionada, como k-NN, árvores de decisão, SVM e outros. Esses algoritmos geralmente são treinados para reconhecer as classes dos conceitos a partir de padrões linguísticos, *topic signatures* ou relações semânticas da WordNet. Fortuna *et al.* [64], por exemplo, propõem classificadores SVM treinados com palavras-chave para categorizar os novos conceitos em uma hierarquia ontológica.

2.5.3 Ambientes para construção semiautomática de ontologias

Existem muitos trabalhos na literatura sobre a construção e enriquecimento de estruturas ontológicas a partir de textos. No entanto, mesmo com o desenvolvimento e evolução de ferramentas de apoio, a construção de tais estruturas é ainda um processo que exige um grande

esforço humano. Em geral, as ontologias são construídas através de processos manuais ou semiautomáticos. Esses processos, além de tediosos e trabalhosos, exigem tempo e requerem manutenção.

Embora a aquisição totalmente automática de conhecimento ainda seja um desafio, vários ambientes têm surgido como o propósito de minimizar o esforço humano no processo. A seguir são comentados alguns trabalhos relacionados a esse tema, tais como: ASIUM, Text2Onto, OntoLearn e OntoLT.

O propósito da ferramenta ASIUM (Acquisition of Semantic knowledge Using Machine Learning Methods) [58] é auxiliar na construção de ontologias para a língua francesa. Utiliza agrupamento conceitual e hierárquico para gerar a taxonomia. Os conceitos são formados pelos termos que aparecem como objetos indiretos de verbos. O agrupamento conceitual é baseado em verbos e em preposições. A validação da ontologia é feita pelo próprio ontologista.

A ferramenta OntoLearn [212] é adequada para enriquecer ontologias de domínio com conceitos e relações. Faz uso de técnicas de processamento linguístico e estatístico para filtrar os termos da língua inglesa. A base lexical WordNet é utilizada para a interpretação semântica dos termos, especialmente multitermos. A ferramenta usa também a WordNet e métodos de aprendizagem indutiva baseados em regras para extrair as relações entre os conceitos. A validação é feita igualmente pelo especialista do domínio.

O ambiente Text2Onto [43] foi construído a partir do Text-to-Onto [133]. Uma das principais diferenças do Text2Onto em relação ao seu antecessor é que, quando identificada uma mudança no *corpus,* não é necessário reprocessá-lo, o que permite ao usuário rastrear as mudanças de evolução da ontologia. O ambiente utiliza tanto abordagens linguísticas quanto técnicas de aprendizagem de máquina, e é baseado no *framework* GATE[12] . O Text2Onto suporta aprendizado de ontologias a partir de textos em inglês, espanhol e alemão.

O OntoLT *[30]* é um *plug-in* para a ferramenta de desenvolvimento de ontologias Protégé, que suporta a extração e extensão interativa de ontologias a partir de textos em alemão e inglês. A abordagem do OntoLT provê um conjunto de regras de mapeamento que associam funcionalmente entidades linguísticas de coleções de textos anotados com conceitos e atributos candidatos, sob a supervisão do usuário.

Já para a língua portuguesa, tem aumentado os esforços quanto à disponibilização de ambientes para a construção de ontologias. São exemplos desses esforços o ONTOLP desenvolvido por Ribeiro [180] e o ambiente de Baségio [13]. Ambos são abordagens semiautomáticas baseadas em padrões linguísticos e medidas estatísticas.

2.5.4 Dificuldades inerentes ao processo de aprendizagem

Ao longo do processo de aprendizagem de ontologia, existem dificuldades e também desafios quanto à extração de informações relevantes a partir de textos e à organização dessas informações de forma coerente a fim de construir uma estrutura conceitual de qualidade.

A preferência dos pesquisadores por abordagens de natureza híbrida, combinando diferentes técnicas e recursos, se justifica também por ser uma forma de minimizar os problemas inerentes ao processo de aprendizagem. Cimiano *et al.* em [41], por exemplo, destacam três problemas que independem do método de agrupamento conceitual a ser usado e que podem impactar na estrutura ontológica resultante da aprendizagem a partir de textos: as relações de dependências escolhidas para formar os conceitos podem não ser corretamente identificadas nos textos, pois as ferramentas usadas para este fim podem cometer erros; nem todas as dependências identificadas são, de fato, interessantes para distinguir os objetos que formarão os conceitos; e as informações obtidas a partir de textos não serão completas pois, por maior que seja o *corpus* usado, nunca será grande o suficiente para conter todas as possíveis dependências.

[12] http://gate.ac.uk/

Os dois primeiros problemas geralmente são enfrentados combinando técnicas linguísticas, que são responsáveis pela identificação das dependências escolhidas, com técnicas estatísticas, que servem para valorizar as dependências e assim viabializar o descarte daquelas cuja importância é considerada menor. Como o peso dessas relações é comumente estimado com base em suas frequências no *corpus*, supõe-se normalmente que os erros ou dependências menos significativas estão embutidos nas relações menos frequentes. Já o terceiro problema tem sido contornado a partir do uso de diferentes *corpora*, extraídos principalmente da *web*, e recursos linguísticos como tesauros e bases lexicais. A WordNet (Seção 4.6) é frequentemente usada para definir os conceitos ou mesmo enriquecê-los, provendo, por exemplo, sinônimos.

É importante mencionar que, em razão da interpretação não consensual por parte dos pesquisadores quanto à definição de ontologia e, principalmente, aos problemas inerentes à aprendizagem de ontologias a partir de textos, como os três descritos anteriormente, a maioria dos trabalhos encontrados não atinge a totalidade das tarefas descritas por Cimiano [38].

Por exemplo, em muitos trabalhos, a aprendizagem de relações não taxonômicas, incluindo a definição automática de seus rótulos, é pouco abordada [187]. Maedche e Staab em [134] comentam que, dentre as três subtarefas essenciais à aprendizagem de ontologias a partir de texto, que são: extração de conceitos, extração da taxonomias e extração de relações não taxonômicas, a última é considerada a mais difícil. Kavalec e Svatek em [104] mencionam que mesmo para os ontologistas essa tarefa não é trivial, visto que são possíveis várias relações entre instâncias de conceitos mais gerais. Segundo Weichselbraun *et al.* [216] essa dificuldade manual de definir e associar rótulos a relações não taxonômicas tem impactado o processo de construção de ontologias e restringindo a aplicabilidade das mesmas em ambientes mais dinâmicos.

Outro aspecto pouco explorado são os axiomas. Aqueles que tentam realizar essa tarefa automaticamente, geram axiomas muito simples [213]. Trabalhos mais recentes como o de Blanco e Dan [22] têm utilizado, para este fim, papéis semânticos. Blanco e Dan exploram as primitivas semânticas nas relações estabelecidas pelos papéis para caracterizar e restringir propriedades.

Observamos que, em geral, os trabalhos limitam-se às tarefas que vão até a hierarquia de conceitos. Isso provavelmente está relacionado ao fato dessas tarefas já estarem mais consolidadas e, portanto, produzirem de forma automática, ou semiautomática com menor esforço manual, estruturas conceituais de melhor qualidade.

Percebemos também que os trabalhos que mais avançaram nessa área são geralmente voltados à Língua Inglesa. Há resultados significativos, também, para o alemão e o francês. Isso é consequência, possivelmente, da riqueza em recursos linguísticos existentes para tais línguas, diferente do que acontece com a Língua Portuguesa, para a qual apenas estão iniciando os esforços nesse sentido.

Apesar dos avanços da área, notamos que ainda há muito a ser feito. Independentemente da língua, percebemos também que a avaliação das estruturas conceituais resultantes desse processo de aprendizagem é um problema em aberto [222]. As ontologias de referência existentes dificilmente contemplam a totalidade dos elementos desejáveis em uma estrutura conceitual. É comum a ausência, por exemplo, em tais ontologias, de relações não taxonômicas e de axiomas. Além disso, a ausência de métodos formais e comuns de avaliação, dificulta a comparação entre as diferentes abordagens propostas. Mesmo em avaliações realizadas por especialistas humanos, as métricas usadas para avaliação pelos pesquisadores nem sempre coincidem. Dada a importância do processo de avaliação de ontologias, esse é o assunto da próxima seção.

2.6 Avaliação de ontologias

Apesar da diversidade de trabalhos publicados sobre a esse tema, métodos formais e padrões de avaliação para sistemas de aprendizagem de ontologias ainda são um problema em aberto.

De acordo com Obsrt *et al.* em [154], há muitos critérios que podem ser usados para se avaliar uma ontologia, tais como: critérios quanto à cobertura do domínio (riqueza, complexidade e granularidade), critérios quanto ao seu desenvolvimento (casos tratados, cenários, requisitos, aplicações e dados de origem) e quanto a propriedades formais relativas à consistência e completude. Essa avaliação, segundo os autores, inclui aspectos de verificação e validação que podem ser estimados a partir de medidas estruturais, funcionais e de usabilidade [73]:

- Estruturais: essas medidas analisam as ontologias enquanto grafos, geralmente de forma mais quantitativa, sem necessariamente medir aspectos semânticos ou de conteúdo. No entanto, em conjunto com outras medidas, podem ser usadas também para qualificar ou classificar estruturas ontológicas. Provêem informações, que podem ser calculadas automaticamente, quanto à profundidade, amplitude, densidade, modularidade, etc. Cimiano em [38], por exemplo, descreve as ontologias usadas em seus experimentos através de medidas estruturais como: número de nodos (conceitos), número de nodos-folha, altura média dos nodos, altura máxima dos nodos, número máximo de nodos-filho e média de nodos-filho presentes na ontologia.

- Funcionais: essas medidas têm foco em aspectos semânticos, procuram avaliar o modelo conceitual caracterizado pela ontologia. Objetivam estimar a distância entre o modelo da estrutura ontológica construída e o modelo pretendido (Figura 2.1). Essa forma de avaliação é complexa e tem diferentes abordagens. Envolve desde avaliações manuais, realizadas por especialistas humanos no domínio da ontologia, até avaliações automatizadas e baseadas em tarefas, que medem a qualidade do modelo a partir do desempenho na resolução de algum problema. Há também abordagens baseadas em casamento (*matching*) de ontologias, no qual o modelo construído é comparado com algum modelo de referência já existente. Para esse fim, as medidas mais usuais são precisão e *recall* [190].

- De usabilidade: essas medidas avaliam os metadados sobre a ontologia e seus elementos. Medem aspectos como a facilidade de acesso às instruções referentes à forma de utilizar a ontologia, controle de versões, compatibilidade, interface com o usuário, etc. Medidas de usabilidade podem ser encontradas em [74].

Nas subseções seguintes, abordamos medidas de caráter estrutural e funcional. Nos deteremos apenas nesses dois tipos, pois acreditamos que medidas referentes à usabilidade vão além do escopo deste trabalho.

Na Seção 2.6.1, apresentamos as métricas estruturais de avaliação utilizadas pelo sistema AKTiveRank [2]. As métricas desse sistema são calculadas de forma relativa. Elas utilizam, como parâmetros, termos informados pelo usuário em uma pesquisa. Desta forma, o sistema é capaz de medir o quanto uma ontologia é abrangente, densa, similar semanticamente e relevante para um conjunto de termos. Escolhemos essas métricas justamente por serem relativas, e, portanto, permitem análises estruturais mais pontuais sobre determinados elementos de uma ontologia.

Para que pudéssemos medir a aplicabilidade de nossa proposta, incluímos em nossos estudos, também, abordagens funcionais de avaliação. Na Seção 2.6.2, descrevemos abordagens funcionais comumente utilizadas e os diferentes aspectos (níveis) ontológicos que tais abordagens são capazes de analisar.

2.6.1 Métricas estruturais do sistema AKTiveRank

O propósito das métricas estruturais é prover informações quantitativas sobre uma ontologia. As métricas propostas na literatura, geralmente, são extensões ou adaptações de medidas conhecidas e buscam atender às peculiaridades das ontologias e de suas aplicações. Como tais

medidas são objetivas e não requerem esforço manual, elas se tornaram um recurso interessante para a organização das ontologias em *ranks*. O sistema AKTiveRank, descrito por Alani e Brewster em [2], é um exemplo. Ele possui um conjunto de métricas que permite escolher, dentre várias ontologias, a mais relevante para uma determinada pesquisa em um sistema de recuperação de informações.

O sistema possui quatro medidas estruturais: *class match*, densidade, similaridade semântica e *betweenness*. Como já mencionado, elas avaliam a ontologia de forma relativa, considerando os termos informados na pesquisa. De acordo com Alani e Brewster, elas podem se definidas como descrito em [2]:

- A medida *class match* (CMM) avalia a cobertura da ontologia para um conjunto de termos previamente informado. São procurados na ontologia aqueles rótulos de conceitos que casam exatamente ou parcialmente com tais termos. Considerando, um conjunto de termos T informados pelo usuário e os conceitos $C(o)$ de uma ontologia o, a medida CMM pode ser definida como a soma dos casamentos exatos E e parciais P existentes entre os rótulos dos conceitos em o e os termos em T. A medida CMM, apresentada na Equação 2.1, permite ainda que se definam pesos para os casamentos. Alani e Brewster usaram, em seus experimentos, $\alpha = 0.6$ e $\beta = 0.4$ como pesos dos casamentos exatos e parciais, respectivamente.

$$CMM(o,T) = \alpha E(o,T) + \beta P(o,T), \text{ onde:} \tag{2.1}$$

$$E(o,T) = \sum_{c \in C(o)} \sum_{t \in T} I(c,t), \text{ para } I(c,t) = \begin{cases} 1 : & se\ rótulo(c) = t \\ 0 : & se\ rótulo(c) \neq t \end{cases}$$

$$P(o,T) = \sum_{c \in C(o)} \sum_{t \in T} J(c,t), \text{ para } J(c,t) = \begin{cases} 1 : & se\ rótulo(c)\ contém\ t \\ 0 : & se\ rótulo(c)\ não\ contém\ t \end{cases}$$

- A medida densidade (DEM) tem como meta avaliar o nível de detalhamento dos conceitos em uma ontologia. Para isso, ela calcula a quantidade de relações não taxonômicas, subclasses, superclasses e classes-irmãs desses conceitos. A medida considera apenas os conceitos, denotados por $C_m(o)$, que casam exatamente ou parcialmente com os termos em T. A Equação 2.2 descreve a medida DEM, na qual o conjunto S_1 corresponde às relações não taxonômicas; S_2, às superclasses; S_3, às subclasses e S_4, às classes-irmãs dos conceitos $c \in C_m(o)$, considerando que $C_m(o) \subseteq C(o)$. A medida admite ainda que sejam definidos pesos w_i para cada conjunto S, o que permite avaliar de forma diferenciada os elementos relacionados aos conceitos.

$$DEM(o) = \frac{1}{n}\sum_{i=1}^{n} dem(c), \text{ onde } n = |C_m(o)| \text{ e } dem(c) = \sum_{i=1}^{4} w_i |S_i| \tag{2.2}$$

- A medida de similaridade semântica (SSM) calcula quão próximos estão, na ontologia o, os conceitos que casam exatamente ou parcialmente com os termos em T. Segundo Alani e Brewster, em um sistema de recuperação, quanto mais longe estão, uns dos outros, os conceitos relacionados aos termos pesquisados, maiores são as chances de a ontologia não representar o conhecimento procurado de uma forma coerente e compacta [2]. Usualmente, a similaridade entre dois conceitos é calculada considerando os caminhos existentes entre esses conceitos em uma estrutura conceitual. Existem várias medidas na literatura que calculam a distância semântica entre conceitos [28], tais como as medidas

de Wu e Palmer [224] e de Leacock e Chodorow [122], que podem ser usadas no cálculo da medida SSM. A Equação 2.3 descreve a medida SSM, na qual P corresponde ao conjunto de todos os caminhos existentes entre os conceitos c_i e c_j, e $c_i \overset{p}{\rightsquigarrow} c_j$ representa um caminho $p \in P$, sendo que $c_i, c_j \in C_m(o)$.

$$SSM(o) = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} ss(c_i, c_j), \text{ onde } ss(c_i, c_j) = \begin{cases} \frac{1}{length(min_{p \in P}(c_i \overset{p}{\rightsquigarrow} c_j))} : & se\ i \neq j \\ 0 : & se\ i = j \end{cases} \quad (2.3)$$

- A medida *betweenness* (BEM) é baseada no trabalho de Freeman [67], que estuda métricas para determinar o nodo central em um grafo. Um nodo é considerado central quando a quantidade de caminhos mais curtos que passam por este nodo é alta. A medida BEM tem como propósito determinar essa quantidade de caminhos mais curtos para os conceitos de uma ontologia. Alani e Brewster acreditam que, quando a medida BEM de um conceito é a mais alta na ontologia, esse conceito corresponde ao elemento central de tal estrutura. A medida BEM é apresentada na Equação 2.4, na qual σ_{c_i,c_j} é o comprimento do menor caminho entre os conceitos c_i e c_j, e $\sigma_{c_i,c_j}(c)$ é quantidade de caminhos curtos entre c_i e c_j que passam pelo conceito c, para $c_i, c_j, c \in C_m(o)$.

$$BEM(o) = \frac{1}{n}\sum_{k=1}^{n} bem(c_k), \text{ onde } \quad bem(c_k) = \sum_{c_i \neq c_j \neq c \in C_m(o)} \frac{\sigma_{c_i,c_j}(c)}{\sigma_{c_i,c_j}} \quad (2.4)$$

Tais medidas, uma vez definidas, são reunidas em um único escore, o qual pode ser usado posteriormente para estabelecer a classificação da ontologia. Considerando o $M_1(o)$ como o valor da medida CMM; $M_2(o)$, o valor de DEM; $M_3(o)$, o valor de SSM e $M_4(o)$, o valor de BEM para a ontologia o; e ainda w como os pesos dessas medidas e O, o conjunto das ontologias sob avaliação, o Escore da ontologia $o \in O$ pode ser definido como apresentado na Equação 2.5. Cabe ressaltar que, para viabilizar a organização das ontologias em um *rank*, as medidas são normalizadas para o intervalo $[0; 1]$.

$$Escore(o \in O) = \sum_{i=1}^{4} w_i \frac{M_i(o)}{max_{1 \leq j \leq |O|}(M_i(o_j))}, \text{ onde } o_j \in O \quad (2.5)$$

Outros estudos e propostas referentes a medidas estruturais podem ser encontrados em [68, 74, 204]. A seguir apresentamos algumas abordagens funcionais, e os diferentes aspectos (níveis) de uma ontologia que tais abordagens podem avaliar.

2.6.2 Abordagens funcionais e níveis de avaliação

De acordo com Brank *et al.* [25] as abordagens de avaliação mais comuns entre os pesquisadores possuem um caráter mais funcional e se enquadram em uma das quatro categorias mencionadas a seguir:

- abordagens baseadas na comparação da estrutura ontológica com uma ontologia de referência (*gold standard*) [41];
- abordagens baseadas no uso da ontologia em uma aplicação e avaliação dos resultados desta última [152];
- abordagens envolvendo comparações com recursos de dados, como um *corpus*, sobre o domínio coberto pela ontologia [27, 142] e
- abordagens em que a avaliação é realizada por juízes humanos os quais tentam estimar o quanto a ontologia atinge critérios pré-definidos, padrões, requisitos, etc [215, 216].

As abordagens que avaliam as estruturas ontológicas a partir de aplicações ou de juízes humanos têm vantagens sobre as demais [49]. Para o uso em uma aplicação, não é necessário avaliar a estrutura ontológica em si. Logo, a avaliação se torna mais objetiva, considerando-se apenas os resultados gerados pela aplicação. Por outro lado, a avaliação também é mais específica; as conclusões acerca dos resultados não podem ser generalizadas para outras aplicações [25].

No caso do uso de especialistas de domínio no processo avaliativo, a vantagem está no olhar humano o qual permite uma análise mais subjetiva, levando em conta aspectos semânticos que abordagens automáticas são incapazes de considerar. Apesar disso, as abordagens automáticas baseadas em ontologias de referência e em *corpus* são as únicas factíveis quando se trata de avaliações em grande escala e de comparações entre múltiplas abordagens de aprendizagem de ontologia [49].

Em abordagens automáticas de avaliação é mais usual que o processo avaliativo seja realizado em níveis [25]. Como uma ontologia é uma estrutura de conhecimento complexa, é mais prático analisá-la sob diferentes aspectos, observando separadamente seus elementos e organização, do que examiná-la como um todo. Cada nível de avaliação considera um determinado aspecto da estrutura conceitual. De acordo com Brank *et al.* [25], uma ontologia pode ser avaliada em níveis:

- léxico: este nível analisa o vocabulário (conceitos, instâncias, etc) usado na estrutura ontológica. Para isso, costumam ser usadas medidas de similaridades entre cadeias de caracteres. É usual também a realização de comparativos entre o vocabulário da estrutura e termos de um *corpus* ou ontologia de referência de mesmo domínio.

- taxonômico: este nível analisa as relações hierárquicas entre os conceitos da estrutura ontológica. Geralmente, as relações taxonômicas da estrutura são comparadas às relações também taxonômicas de uma ontologia de referência de mesmo domínio. São usadas, nessa análise comparativa, medidas como precisão e abrangência.

- não taxonômico: este nível analisa as relações transversais da estrutura conceitual. Como as relações taxonômicas, as tranversais podem ser avaliadas a partir de uma ontologia de referência. No entanto, são raras as ontologias de referência que contêm relações não taxonômicas [187], o que tem dificultado a aplicação de tal abordagem. Por esta razão, vários pesquisadores têm optado por avaliações a partir da análise de especialistas humanos. Apesar disso, têm surgido novas propostas incluindo medidas intrínsecas de similaridade baseadas na WordNet [187] e em conteúdo de informação [191] para avaliar tais relações.

- contextual: este nível não analisa a ontologia em si, mas o contexto em que será usada. Avalia de forma prática a adequação da estrutura ontológia para uma aplicação em particular. A análise é realizada a partir dos resultados gerados pela aplicação.

- sintático: este nível de avaliação é especialmente interessante para ontologias desenvolvidas manualmente. Analisa aspectos referentes à linguagem, tais como uso correto de palavras-chave referentes ao domínio e a presença de *loops* nas definições dos conceitos da ontologia.

- de projeto: este nível analisa a arquitetura da estrutura conceitual e é aplicado normalmente quando a ontologia é construída de forma manual. Avalia a organização da estrutura conceitual, sua facilidade de manutenção e, ainda, a qualidade de sua documentação para uso em aplicações.

A Tabela 2.1 relaciona as abordagens mais usuais, anteriormente citadas, aos níveis de avaliação mencionados. Como se pode observar, os níveis mais amplamente abordados são o léxico, o taxonômico e o não taxonômico.

Tabela 2.1 – Níveis ontológicos analisados pelas abordagens mais usuais para avaliação de estruturas conceituais (adaptado de [25])

	Abordagem Baseada em			
nível	ontologia de referência	aplicação	dados	juízes humanos
léxico	x	x	x	x
taxonômico	x	x	x	x
não taxonômico	x	x	x	x
contextual		x		x
sintático	x			x
de projeto				x

Nas subseções seguintes, comentamos os quatro primeiros níveis de avaliação: léxico, taxonômico, não taxonômico e contextual. Os níveis sintático e de projeto não são detalhados neste documento por serem mais indicados para abordagens manuais.

2.6.2.1 Nível léxico

Como já mencionado, a avaliação em nível léxico consiste em analisar os rótulos usados para identificar os conceitos (vocabulário) de uma ontologia. O vocabulário da ontologia sob análise, geralmente, é comparado a um outro vocabulário, sendo que este último é composto de rótulos de conceitos que podem ser provenientes de uma ontologia de referência, gerados estatisticamente a partir de um *corpus* ou preparados por especialistas do domínio [25].

Em abordagens para aprendizagem de ontologias, é muito comum que essa comparação seja realizada em relação a uma ontologia de referência [49]. É usual também a aplicação de medidas como precisão e abrangência nesse processo comparativo. Dellschaft e Staab em [49] chamam essas medidas de precisão léxica (PL) e abrangência léxica (AL) . Essas medidas permitem comparar o vocabulário de uma ontologia gerada automaticamente O_A ao de uma ontologia de referência O_R. As Equações 2.6 e 2.7 apresentam, respectivamente, as medidas PL e AL, onde C_A corresponde ao conjunto de rótulos de conceitos de O_A e C_R, aos de O_R.

$$PL(O_A, O_R) = \frac{|C_A \cap C_R|}{|C_A|} \qquad (2.6) \qquad\qquad AL(O_A, O_R) = \frac{|C_A \cap C_R|}{|C_R|} \qquad (2.7)$$

Um dos problemas dessa forma de avaliação, no entanto, é a não correspondência perfeita entre os vocabulários das ontologias que estão sendo comparadas. O problema de "casamento de cadeias de caracteres" tem sido minimizado através do uso de funções *strings*, que permitem verificar se um rótulo está contido em outro (é seu *substring*), ou por meio de medidas de distância entre cadeias de caracteres, como a de Levenshtein [124]. Esse é um problema comum, em avaliação léxica, quando, por exemplo, a ontologia de referência é a WordNet e a ontologia sob análise contém muitos conceitos cujos rótulos são nomes compostos.

Outro problema ocorre quando os conceitos não possuem rótulos ou estes são considerados inadequados. Este problema é mais frequente em abordagens automáticas, que utilizam algoritmos de agrupamento para identificar os conceitos, como o método FCA. Nesses casos, a abordagem mais comum é a aplicação de medidas de similaridade entre os elementos que compõem os conceitos. Um exemplo desse tipo de similaridade é apresentado na Seção 3.7.2, que descreve a medida Sim, usada para comparar os conceitos gerados pelo FCA. A medida Sim tem como base a medida ics (information content similarity) que utiliza a WordNet e um *corpus* para estimar a similaridade semântica entre termos.

Cabe destacar ainda que as medidas PL e AL não levam em consideração o aspecto polissêmico[13] dos rótulos que identificam os conceitos em uma ontologia, o que obviamente pode

[13]Uma palavra é dita polissêmica quando pode assumir significados diferentes, conforme o contexto em que

prejudicar a validade e interpretação de seus resultados.

2.6.2.2 Nível taxonômico

Também no caso das relações hierárquicas é comum que a avaliação seja realizada com base em ontologias de referência. O objetivo, neste caso, é analisar o quanto a ontologia sob avaliação está estruturalmente alinhada a uma ontologia de referência [25].

Dellschaft e Staab em [49] propõem medidas para este nível de avaliação, que são aplicáveis em aprendizagem de ontologias. As medidas propostas pelos autores são extensões do trabalho de Maedche e Staab [130] e permitem a análise das relações de forma local (analisando conceitos) e global (analisando a taxonomia por completo) quanto à precisão e abrangência. Segundo Dellschaft e Staab [49], a precisão taxonômica local (pt) analisa a similaridade entre dois conceitos de ontologias distintas. Essa medida calcula a proporção de características extraídas de suas respectivas estruturas conceituais que são compartilhadas pelos conceitos. A Equação 2.8 define a medida pt, onde O_A corresponde à ontologia construída automaticamente, O_R indica a ontologia de referência, $carac$ é a operação que extrai as características de um conceito e, c_A e c_R são conceitos das ontologias O_A e O_R, respectivamente.

$$pt_{carac}(c_A, c_R, O_A, O_R) = \frac{|carac(c_A, O_A) \cap carac(c_R, O_R)|}{|carac(c_A, O_A)|} \tag{2.8}$$

Em nossa pesquisa [25, 222], observamos que as operações mais usuais para extração de características taxonômicas em processos avaliativos são semantic cotopy (sc) [130] e sua variação, common semantic cotopy (csc) [49]. A operação sc estabelece, para um conceito c de uma determinada ontologia, o conjunto de conceitos que se relacionam hierarquicamente com c, ou seja, seus super ou subconceitos. Já csc, ao construir esse conjunto, considera apenas os conceitos comuns às duas ontologias sob comparação. Por esta razão, a csc é menos dependente lexicamente do que sc, visto que desconsidera conceitos que não existem em ambas ontologias [49]. As Equações 2.9 e 2.10 definem, respectivamente, as operações sc e csc, ond C_A é o conjunto de conceitos de O_A, C_R, o de O_R e C, o de qualquer ontologia O [49].

$$sc(c, O) = \{c_i | c_i \in C \wedge (c_i \leq c \vee c \leq c_i\} \tag{2.9}$$

$$csc(c, O_A, O_R) = \{c_i | c_i \in C_A \cap C_R \wedge (c_i \leq c \vee c \leq c_i\} \tag{2.10}$$

Decidida a operação de extração de características, que será usada para calcular a precisão local, pode-se então definir a precisão global, que permitirá obter a precisão da taxonomia de O_A em relação a O_R. A precisão global, chamada de precisão taxonômica (PT), é apresentada na Equação 2.11. Cabe ressaltar que, quando um conceito c não é comum às duas ontologias, a precisão local é estimada. Nesse caso, ela é definida como a maior precisão local encontrada entre os valores precisão local calculados para todos os conceitos c' em O_R em relação a c [49].

$$PT_{carac}(O_A, O_R) = \frac{1}{|C_A|} \sum_{c \in C_A} \begin{cases} pt_{carac}(c, c, O_A, O_R) & se \;\; c \in C_R \\ max_{c' \in C_R} pt_{carac}(c, c', O_A, O_R) & se \;\; c \notin C_R \end{cases} \tag{2.11}$$

A partir da medida de precisão PT pode-se calcular também a abrangência AT [49] e a medida $F_\beta T - score$ [74] taxonômicas, as quais são apresentadas, respectivamente, nas Equações 2.12 e 2.13.

é aplicada [99]. A relação de polissemia é aquela em que os itens possuem a mesma forma (grafia) e diferentes significados [162]. A granularidade da diferença semântica pode variar. Pode ser bem expressiva tal como "verde" para indicar uma cor e "verde" para indicar que algo não está maduro. Mas também pode ter significados relacionados, como a palavra banco quando aplicada no sentido de “repositório”, pode ser um “banco de sangue”, um “banco de dados”, um “banco de células”, etc [229]. O mesmo se aplica a verbos, como por exemplo: "pintar um quadro" e "pintar uma parede".

$$AT_{carac}(O_A, O_R) = PT_{carac}(O_R, O_A) \tag{2.12}$$

$$F_\beta T_{carac}(O_A, O_R) = \frac{(1+\beta^2) \times PT_{carac}(O_R, O_A) \times AT_{carac}(O_A, O_R)}{\beta^2 \times PT_{carac}(O_R, O_A) + AT_{carac}(O_A, O_R)} \tag{2.13}$$

Dellschaft e Staab [49] descrevem ainda a medida $F_\beta T'$ (Equação 2.14) que permite avaliar a qualidade da ontologia tanto em nível léxico quanto taxônomico, incluindo na equação a medida de precisão léxica (Equação 2.6). Em [49], os autores usaram $\beta = 1$ para definir as medidas $F_\beta T$ e $F_\beta T'$ [49].

$$F_\beta T'(O_A, O_R) = \frac{(1+\beta^2) \times PL(O_R, O_A) \times FT_{carac}(O_A, O_R)}{\beta^2 \times PL(O_R, O_A) + FT_{carac}(O_A, O_R)} \tag{2.14}$$

Outra medida comentada é a Taxonomic Overlap (TO) que mede a sobreposição média de relações hierárquicas de O_A em relação a O_R. A equação para o cálculo de TO, apresentada em 2.15, é semelhante à de PT, pois também utiliza uma medida local de sobreposição, denotada por to_{sc}. A medida to_{sc} (Equação 2.16) é baseada na medida sc.

$$TO(O_A, O_R) = \frac{1}{|C_A|} \sum_{c \in C_A} \begin{cases} to_{sc}(c, c, O_A, O_R) & se \;\; c \in C_R \\ max_{c' \in C_R} to_{sc}(c, c', O_A, O_R) & se \;\; c \notin C_R \end{cases} \tag{2.15}$$

$$to_{sc}(c_A, c_R, O_A, O_R) = \frac{|sc(c_A, O_A) \cap sc(c_R, O_R)|}{|sc(c_A, O_A) \cup sc(c_R, O_R)|} \tag{2.16}$$

Alguns autores usam a WordNet como ontologia de referência. No entanto, tal abordagem para avaliação de ontologias de domínio é criticada, pois as relações de hiponímia e hiperonímia da WordNet podem não refletir por completo as relações hierárquicas de conceitos em domínios específicos [187]. Uma alternativa tem sido validar as relações hierárquicas a partir das categorias da Wikipédia. O trabalho de Yu *et al.* [228] é um exemplo. Nesse trabalho, os autores comparam a estrutura de navegação proporcionada pelas relações entre os conceitos da ontologia sob análise com a navegação permitida pelas categorias da Wikipédia.

É importante destacar que essas foram algumas das poucas medidas que encontramos, em nível taxonômico, aplicadas também a estruturas conceituais geradas a partir do FCA, enquanto método de agrupamento conceitual (como em [41]). Esse aspecto foi determinante para a inclusão dessas medidas em nosso estudo.

2.6.2.3 Nível não taxonômico

Embora existam diferentes abordagens para avaliação de relações transversais, como mostra a Tabela 2.1, há várias dificuldades inerentes a esse processo. As dificuldades, no entanto, não são só quanto às métricas a serem utilizadas, mas também quanto aos recursos disponíveis para esse tipo de avaliação.

Brank *et al.* [25] argumentam que as medidas como precisão e abrangência usadas em nível taxonômico podem ser adaptadas para analisar relações não taxonômicas [11, 104, 216]. No entanto, quando a abordagem de avaliação escolhida é por meio de uma ontologia de referência, a dificuldade principal não reside nas métricas, mas na ontologia que será usada para esse fim. Segundo Sánchez e Moreno [187] são raras as ontologias de referência que contêm relações não taxonômicas. A ausência dessas relações é decorrente, principalmente, da dificuldade do ontologista em defini-las dado o volume e as variações de relacionamentos possíveis entre conceitos [216]. Isso prejudica a avaliação, pois a ausência de uma relação não significa, necessariamente, que ela é incorreta, e ainda penaliza índices como a precisão [104].

Por esta razão, têm surgido novas propostas incluindo medidas intrínsecas de similaridade comumente baseadas na WordNet [187] e em conteúdo de informação [191] para avaliar tais relações. No entanto, os próprios autores consideram ainda suas medidas incipientes [131, 187].

A dificuldade relativa ao uso da WordNet, nessa tarefa de avaliação, está no fato de esse recurso conter relações mais gerais e não de domínio. Desta forma, o mais usual é que os resultados seja analisados por especialistas humanos [131, 187, 192, 215].

2.6.2.4 Nível contextual

Nesse nível de avaliação, o mais comum é que a ontologia seja analisada a partir do seu desempenho no apoio à realização de uma tarefa. Nesse caso, a avaliação é extrínseca, são utilizadas métricas de avaliação relativas à tarefa e não se analisa diretamente a ontologia enquanto estrutura de representação de conhecimento.

Segundo Brank *et al.* [25], a avaliação nesse nível é mais objetiva quando comparada às dos níveis já comentados, mas tem algumas desvantagens. A avaliação em nível contextual permite se observar o quanto uma estrutura ontológica é boa ou ruim para uma tarefa específica, no entanto torna difícil a generalização dessa informação no que se refere a outras tarefas. Além disso, se a ontologia for dependente de um componente da aplicação, a análise dos resultados poderá ser prejudicada. Pode ser difícil medir separadamente o desempenho do componente e a contribuição, propriamente dita, da ontologia. E ainda, no caso de várias ontologias, pode ser difícil também decidir qual a melhor para uma aplicação, se esta aplicação for flexível o suficiente para desempenhar de forma satisfatória com qualquer uma das ontologias sob avaliação.

Obviamente que, pelo fato de as ontologias serem estruturas de representação, suas principais aplicações estão na área de sistemas de informação. Elas têm sido usadas, principalmente, para apoiar tarefas de classificação [23, 152, 223], agrupamento [103, 115, 148, 149] e recuperação de informações [35, 175, 231]. E é, portanto, para esse tipo de tarefa que as avaliações de contexto têm se direcionado.

Em nossa pesquisa, encontramos poucos trabalhos cujo tema central seja o desenvolvimento de metodologias para avaliação de ontologias a partir de tarefas. O trabalho de Netzer *et al.* em [152] é um dos poucos que têm essa proposta. Os autores apresentam um método para avaliar estruturas ontológicas por meio da tarefa de classificação de textos em aplicações de recuperação de informações. Os autores analisam, por exemplo, a cobertura léxica da ontologia com relação aos termos utilizados em consultas e a qualidade de classificação de textos proporcionada pelo uso da ontologia no que tange à recuperação desses textos.

2.7 Considerações sobre este capítulo

A partir do estudo realizado sobre estruturas ontológicas e os métodos usados para extraí-las de textos, percebemos a falta de consenso entre os pesquisadores sob diferentes aspectos, que vão desde a definição do que é uma estrutura ontológica aos métodos e técnicas usados para construí-las.

No que tange à definição, adotaremos a de Gruber em [85] que é amplamente aceita, sendo referenciada por muitos trabalhos da área. No entanto, ao invés do termo "ontologia" usaremos tanto o termo "estrutura ontológica" como "estrutura conceitual" para nos referirmos a esse tipo de entidade, respeitando, portanto, a abrangência da definição de Gruber e, também, indo ao encontro da visão de autores como Breitman *et al.* [26] quanto à classificação de ontologias (Seção 2.2).

Quanto à metodologia, percebe-se algumas linhas gerais de procedimento, como as tarefas enumeradas por Cimiano [38] e apresentadas na Seção 2.5.1 para aprendizagem de estruturas ontológicas a partir de textos. Mas, de uma forma geral, os trabalhos estudados utilizam métodos diversos para realizar essas tarefas, sendo que nem todas as etapas que estudamos são de fato implementadas. Muitas ontologias são apenas taxonomias e não incluem axiomas e nem relações transversais.

Por outro lado, há um caminho de pesquisa recorrente e, por isso, talvez consensual, que se refere ao uso de abordagens híbridas e ao uso de recursos *web*. Muitos trabalhos adotam técnicas lingüísticas, estatísticas e de aprendizagem de máquina para realizar as tarefas. Assim como se apóiam fortemente na WordNet quanto à identificação de conceitos e relações semânticas. Em trabalhos mais recentes, observa-se ainda o uso papéis semânticos tanto para identificar e rotular relações não taxonômicas [11, 193] quanto para definir axiomas [22]. Além disso, já é uma prática o uso de *corpora* construídos a partir de documentos *web* e de textos da Wikipédia.

As abordagens são, em sua ampla maioria, semiautomáticas, o que é justificável dada a complexidade envolvida na construção de estruturas ontológicas a partir de textos. Há uma certa preferência no se refere às avaliações intrínsecas dessas estruturas. As avaliações dessa natureza, que são realizadas automaticamente, em geral têm um caráter mais estrutural. Assim, quando o objetivo é analisar a qualidade semântica da ontologia enquanto estrutura de representação de conhecimento, a avaliação acaba sendo manual. Isso se deve à lacuna que existe ainda quanto a medidas de avaliação de cunho semântico que possam ser aplicadas automaticamente e forneçam informações relevantes e qualitativas sobre a estrutura.

Alternativamente, alguns pesquisadores têm adotado avaliações extrínsecas, de caráter funcional, comparando as estruturas geradas com ontologias de referência [41] ou comparando-as com base no desempenho de aplicações [152]. No entanto, quando a avaliação tem como objetivo qualificar as relações não taxonômicas, a ausência de tais relações em ontologias de referência e de métodos e métricas formais para esse tipo de avaliação têm dificultado tanto a aplicação de abordagens funcionais quanto a comparação dos resultados de pesquisas nessa área.

Já quanto ao aspecto tecnológico, não há ferramentas que suportem todo o processo de aprendizagem e de avaliação de tais estruturas. Faltam também *benchmarks* para a área de aprendizagem de ontologias a partir de textos, para que seja possível, por exemplo, comparar metodologias e assim estabelecer quais métodos e técnicas são mais indicados para cada tarefa, conforme a natureza do domínio do *corpus* [80].

É evidente, também, que os maiores avanços nessa área estão voltados para abordagens relativas à Língua Inglesa. Acreditamos que isso esteja diretamente ligado à riqueza de recursos disponíveis para processá-la. No caso da Língua Portuguesa, o avanço não é tão intenso, possivelmente porque muitos recursos ainda estão em desenvolvimento, como é o caso da WordNet.br [196].

Observamos, por fim, que há ainda muito espaço de investigação na área de aprendizagem de ontologias a partir de texto, cabendo, sem dúvida, novos estudos e propostas, incluindo métodos e abordagens diversas. Sendo este o caso, inclusive, do método Formal Concept Analysis que, embora não seja novo e nem seja de aplicação recente na área, tem despertado o interesse dos pesquisadores e, consequentemente, vem se destacando como uma alternativa de método de agrupamento conceitual na construção de estruturas ontológicas. Como esse método é parte do nosso estudo, ele é abordado no próximo capítulo.

3. FORMAL CONCEPT ANALYSIS

Este capítulo faz uma introdução ao método Formal Concept Analysis (FCA), apresentando sua definição e embasamento matemático. Descrevemos também uma extensão do FCA: o método Relational Concept Analysis (RCA). Segundo Priss [168], o RCA é indicado para a representação de relações não taxonômicas. Posto que nossa pesquisa tem interesse em papéis semânticos os quais estabelecem relações dessa mesma natureza, decidimos investigar tal extensão. Neste capítulo são comentados ainda, algoritmos para gerar estruturas[1] FCA; métodos que procuram reduzir a complexidade computacional desses algoritmos; e medidas de similaridade apropriadas para tais estruturas. Além disso, apresentamos as vantagens e desvantagens do FCA enquanto método de agrupamento conceitual, bem como algumas de suas aplicações na área de aprendizagem de ontologias a partir de textos.

Embora muitas formas de estruturas conceituais sejam tratadas, hoje, de ontologias como foi comentado na Seção 2.2, cabe ressaltar que, em nossos exemplos iniciais, o foco restringe-se mais ao método do que a sua aplicação. É importante esclarecer que nem sempre o que chamamos de "conceito formal" corresponde a um "conceito ontológico" no sentido mais usual em ciência da computação, ou seja, a uma classe. Em alguns exemplos, o "conceito formal" é composto tanto de classes quanto de instâncias. Apesar de os métodos geralmente organizarem as classes nos nodos superiores e as instâncias nos nodos mais inferiores, entendemos que a distinção entre o que é classe e o que é uma instância seja um passo anterior ao uso desses métodos formais. Portanto, essa distinção entre classe e instância é feita apenas nos exemplos de aplicações dos referidos métodos para construção de estruturas ontológicas a partir de textos, visto que para esse fim tal distinção é necessária.

3.1 Origem e características

Formal Concept Analysis (FCA) foi introduzido pelo matemático alemão Rudolf Wille na década de 80 como um método para análise de dados. A estrutura organizacional proposta pelo método, que é baseada na teoria dos reticulados [48], permite a visualização, a investigação e a interpretação dos dados e de suas estruturas, implicações e dependências inerentes [76].

O método tem sido aplicado em diferentes áreas, tais como psicologia, sociologia, antropologia, medicina, biologia, linguística, ciência da computação, matemática e engenharia industrial [221]. Na área de ciência da informação é usado para análise de dados, representação de conhecimento e gestão de informação [171]. Na ciência da computação, o método tem sido utilizado para a construção automática de estruturas conceituais, já que a organização que ele provê para os dados pode ser vista como uma técnica de agrupamento conceitual [38].

O FCA estrutura os dados em unidades, chamadas conceitos formais, e as organiza na forma de um reticulado de conceitos. Os conceitos formais e os reticulados de conceitos são abstrações matemáticas do que, em filosofia, chamamos de conceitos e de hierarquia de conceitos, respectivamente [220].

Do ponto de vista filosófico, segundo Wille em [220], os conceitos pode ser entendidos como unidades básicas do pensamento humano, cuja formação decorre de processos dinâmicos em ambientes sociais e culturais. Os conceitos, ainda nesse sentido, são caracterizados por suas extensões e intensões. As extensões compreendem todos os objetos que pertencem a esses conceitos. Já as intensões expressam os atributos que descrevem as propriedades e os significados de todos os objetos contidos nas extensões dos conceitos. Wille comenta também que os

[1] Ao longo do texto, usamos o termo "estrutura FCA" para nos referir ao reticulado de conceitos resultante da aplicação do método FCA.

conceitos se relacionam de forma hierárquica. Assim, um relacionamento do tipo subconceito-superconceito indica que a extensão do subconceito está contida na extensão do superconceito e ainda que a intensão do subconceito contém a intensão do superconceito.

Formalmente, essa definição filosófica de conceito é mapeada na dualidade conhecida como "conexão Galois", que estabelece relações implícitas entre dois conjuntos parcialmente ordenados, no caso objetos e atributos, de forma que os objetos possam ser descritos através de seus atributos e os atributos, pelos objetos que caracterizam [171].

As definições matemáticas do método são apresentadas na próxima seção.

3.2 Conceitos formais e reticulados de conceitos

O modelo matemático que permite descrever as extensões e intensões dos conceitos formais, inicialmente introduz a noção de contexto formal. Contextos formais são caracterizados pela tripla (G, M, I), onde [38, 171, 220]:

- G é o conjunto formado pelas entidades do domínio, ditas objetos formais;
- M é constituído pelas características dessas entidades, seus atributos formais; e
- I é uma relação binária sobre $G \times M$, chamada relação de incidência, que associa um objeto formal ao seu atributo. Desta forma, a relação gIm pode ser lida como: "o objeto g tem o atributo m".

Em um texto, por exemplo, a partir de dependências sintáticas entre os verbos e seus argumentos, podemos definir contextos formais. A Tabela 3.1 apresenta um subconjunto de relações, entre verbos e seus objetos (diretos e indiretos), extraídas da seção Esportes do *corpus* PLN-BR CATEG (detalhes sobre o *corpus* são apresentados na Seção 5.1.3).

Nessa tabela, o conjunto G de objetos formais é *{campeonato, jogo, ponto, partida}* e o conjunto M de atributos formais é *{perder, marcar, vencer, jogar}*. Já o conjunto I corresponde ao cruzamento dos elementos tabulados (as relações entre G e M), tal como a do objeto *ponto* com os atributos *perder* e *marcar*. Nesse exemplo, os verbos são interpretados como "propriedades" dos objetos. Desta forma, o objeto *ponto* é algo que pode ser *perdido* ou *marcado*.

Tabela 3.1 – Exemplo de contexto formal definido a partir de relações entre verbos e seus argumentos.

	perder	marcar	vencer	jogar
campeonato	x	x	x	x
jogo	x		x	
ponto	x	x		
partida	x			

Para definir os conceitos formais a partir de um contexto formal, são necessários operadores de derivação. Assim, para dois conjuntos arbitrários de objetos e atributos, denotados, respectivamente, por O e A, os operadores O' e A' podem ser definidos como [220]:

$$O' = \{m \in M / gIm \text{ para todo } g \in O\}$$

$$A' = \{g \in G / gIm \text{ para todo } m \in A\}$$

Considerando que O' determina todos os atributos em M compartilhados pelos objetos em O, e A' determina todos os objetos em G que compartilham os atributos em A, um conceito

formal em (G, M, I) é definido pelo par (O, A) se e somente se $O \subseteq G$, $A \subseteq M$, tal que $O' = A$ e $A' = O$ [171, 220].

Os operadores O' e A', portanto, expressam a "conexão Galois", formando conceitos como pares do tipo *(extensão, intensão)*. Da Tabela 3.1 pode-se, então, extrair os conceitos formais:

- $(\{campeonato\}, \{perder, marcar, vencer, jogar\})$,
- $(\{campeonato, jogo\}, \{perder, vencer\})$,
- $(\{campeonato, ponto\}, \{perder, marcar\})$,
- $(\{campeonato, partida\}, \{perder\})$.

Para que os conceitos formais possam ser organizados hierarquicamente, é necessário estabelecer a relação subconceito-superconceito. Matematicamente, esta relação de ordem pode ser definida como [220]:

$$(O_1, A_1) \leq (O_2, A_2) \Longleftrightarrow O_1 \subseteq O_2 (\Longleftrightarrow A_1 \supseteq A_2)$$

A relação subconceito-superconceito, portanto, pode ser de "inclusão da extensão" ($O_1 \subseteq O_2$) ou de "inclusão das intensões, mas em ordem inversa" ($A_1 \supseteq A_2$) [41]. A Figura 3.1 ilustra essas relações de ordem para os conceitos *({campeonato},{perder,marcar,vencer,jogar})* e *({jogo,campeonato}, {perder,vencer})*.

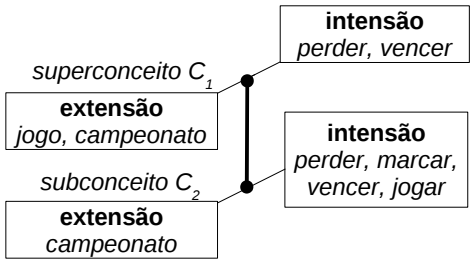


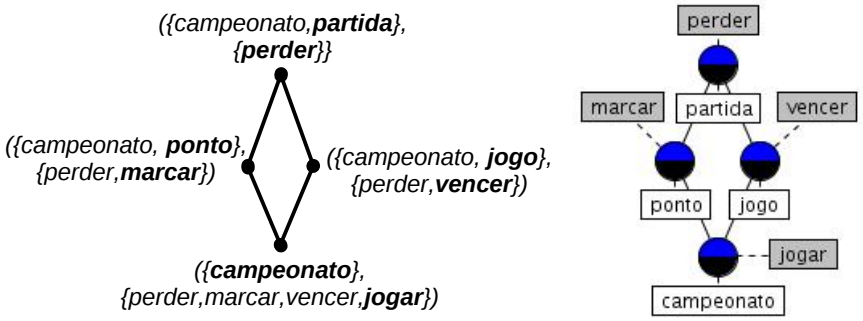
Figura 3.1 – Relação de ordem dos pares (extensão, intensão), adaptado de [171].

O conjunto de todos os conceitos formais de um contexto (G, M, I) juntamente com a relação de ordem formam um reticulado completo, chamado de reticulado de conceitos de (G, M, I) e denotado por $\mathscr{B}(G, M, I)$. Isso significa que, para todo conjunto de conceitos, há um único maior subconceito (o ínfimo) e um único menor superconceito (o supremo) [201].

A Figura 3.2 ilustra duas representações, por diagrama de linhas, para o reticulado de conceitos formado a partir da Tabela 3.1. O diagrama[2] da Figura 3.2b é a representação mais usual e é resultante da técnica de "etiquetagem reduzida" (*reduced labeling*) [76]. Essa técnica é muito útil para reticulados com um grande número de conceitos, pois facilita a visualização da estrutura, omitindo rótulos de objetos e atributos. Por meio desta técnica, em caminhos ascendentes de nodos (conceitos), se um rótulo de objeto pertencer a todos esses nodos, apenas o nodo mais inferior desse caminho exibirá tal rótulo, ficando implícita a sua presença nos seus ascendentes. No caso de atributos, é o inverso. O rótulo aparecerá apenas em um nodo superior, tornando-se sua presença implícita nos seus descendentes.

O contexto formal que originou o reticulado de conceitos apresentado na Figura 3.2 é univalorado. No entanto, em aplicações reais, os contextos costumam ser multivalorados [208]. Tais contextos exigem um pré-processamento antes da geração do reticulado e, por isso, são o assunto da próxima seção.

[2]Este diagrama foi gerado com a ferramenta Concept Explorer 1.3. Mais detalhes sobre a ferramenta podem ser encontrados na Seção 5.3.5.



(a) (b)

Figura 3.2 – Exemplo de representação de um reticulado de conceitos sem e com a técnica de "etiquetagem reduzida"

3.3 Contextos formais multivalorados

Em aplicações reais, o mais comum é descrever os dados por meio de tabelas, nas quais os dados são relacionados por atributos de diferentes tipos (numéricos, ordinais, categóricos, etc). Essas tabelas são definidas formalmente por contextos multivalorados (G, M, W, I), onde [76, 166]:

- G é o conjunto de objetos formais;
- M é o conjunto de atributos formais;
- W é o conjunto de valores de atributos formais e
- I é a relação ternária entre G, M e W ($I \subseteq \{G \times M \times W\}$) que deve satisfazer a condição: "se $(g, m, w) \in I$ e $(g, m, v) \in I$, então $w = v$". A relação (g, m, w) é lida como: "o objeto g tem um atributo m com o valor w". Essa relação pode ser interpretada como uma função parcial de G para W, podendo, portanto, ser reescrita como $m(g) = w$.

A Tabela 3.2 é um exemplo de contexto multivalorado. Ela contém o nome, a data da fundação e o estado de alguns clubes de futebol brasileiros. Nesta tabela, G é o conjunto de clubes *{"Internacional", "Grêmio", "São Paulo", "Palmeiras", "Flamengo", "Vasco"}*, M é conjunto de atributos *{"data de fundação", "estado"}* e W é o conjunto dos valores desses atributos *{"04/04/1909", "15/09/1903", ..., "RS", "SP","RJ"}*.

Tabela 3.2 – Dados de alguns clubes de futebol brasileiros

clubes	data de fundação	estado
Internacional	04/04/1909	RS
Grêmio	15/09/1903	RS
São Paulo	16/12/1935	SP
Palmeiras	26/08/1914	SP
Flamengo	17/11/1895	RJ
Vasco	11/08/1898	RJ

Para que esses dados possam ser organizados em um reticulado de conceitos é necessário transformar esse contexto multivalorado em um contexto univalorado (ou binário) equivalente [208]. Essa transformação é feita mediante o uso de uma escala conceitual e, por esta, razão Ganter e Wille denominam esse processo de *conceptual scaling*. Nesse processo, o mais usual é a aplicação de *plain scales*, principalmente por serem mais simples [76].

O processo consiste em definir uma *plain scale* $S_m = (G_m, M_m, I_m)$ para cada $m \in M$ de um contexto multivalorado (G, M, W, I), onde $m(G) \subseteq G_m$. Cada escala S_m organiza os objetos de uma forma particular e é determinada de acordo com o tipo e as características do atributo [54]. As escalas, em geral, requerem subdivisões, as quais devem ser significativas e selecionadas de forma criteriosa. Por esta razão geralmente são determinadas manualmente [165, 173].

Para representar o atributo ordinal "data de fundação", por exemplo, pode-se optar por diferentes escalas. Utilizamos, em nosso exemplo, a escala $S_{dataDeFundação} = (G_{dataDeFundação}, M_{dataDeFundação}, I_{dataDeFundação})$, onde $G_{dataDeFundação}$ ={*04/04/1909, 15/09/1903, 16/12/1935, 26/08/1914, 17/11/1895, 11/08/1898*}, $M_{dataDeFundação}$ ={*"fundação<=1900", "1900<fundação<=1930", "fundação>1930"*} e $I_{dataDeFundação} \subseteq G_{dataDeFundação} \times M_{dataDeFundação}$ é definida pela Tabela 3.3.

Tabela 3.3 – Relação de incidência $I_{dataDeFundação}$

	"fundação<=1900"	"1900<fundação<=1930"	"fundação>1930"
04/04/1909		x	
15/09/1903		x	
16/12/1935			x
26/08/1914		x	
17/11/1895	x		
11/08/1898	x		

A escolha do conjunto $M_{dataDeFundação}$ foi subjetiva. Poderíamos ter criado uma escala dividida em décadas ou mesmo em intervalos de datas, se os dias e meses fossem julgados relevantes para a representação do domínio.

Já para o atributo categórico "estado" usamos a escala $S_{estado} = (\{"RS","SP","RJ"\}, \{"estado:RS","estado:SP","estado:RJ"\}, é-substring-de)$. No entanto, poderíamos ter usado uma escala mais compacta, considerando, por exemplo, apenas as regiões dos estados.

Definidas as escalas pode-se, então, gerar o contexto univalorado, chamado de contexto formal derivado. Esse contexto é denotado por $(G, \cup_{m \in M} M_m, J)$, onde J é uma relação binária entre G e o conjunto união dos atributos M_m definidos pelas escalas, de tal forma que $gJn \Leftrightarrow wI_mn$ para $(g, m, w) \in I$ e $n \in M_m$. A Tabela 3.4 apresenta o contexto formal derivado a partir das escalas $S_{dataDeFundação}$ e S_{estado}.

A partir do contexto formal derivado (Tabela 3.4) pode-se, então, construir o reticulado de conceitos para os clubes de nosso exemplo (Figura 3.3).

Tabela 3.4 – Contexto formal de alguns clubes de futebol brasileiros após a transformação por *plaine scale*

	clube	fundação<=1900	1900<fundação<=1930	fundação>1930	estado:RS	estado:SP	estado:RJ
Internacional	x		x		x		
Grêmio	x		x		x		
São Paulo	x			x		x	
Palmeiras	x		x			x	
Flamengo	x	x					x
Vasco	x	x					x

Mesmo exigindo um pouco mais de processamento para construir o reticulado de conceitos, quando os contextos são multivalorados, o método FCA mostra-se um recurso interessante para a construção de estruturas conceituais. No entanto, Hacene *et al.*, em [89], afirmam que o método FCA não é suficiente para expressar relações não taxonômicas, que são comuns em estruturas ontológicas.

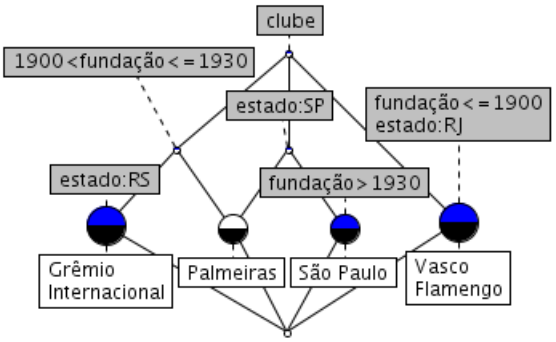


Figura 3.3 – Reticulado de conceitos para clubes de futebol

Segundo Priss em [168], relações semânticas, como papéis semânticos (também foco de nosso estudo) são interpretadas como relações funcionais que não formam ordenações hierárquicas. Portanto, não podem ser representadas pelo método FCA. Para expressar esses tipos de relações nos reticulados de conceitos, seria necessário usar uma extensão do método, conhecida como Relational Concept Analysis.

Cabe ressaltar que, o fato do método FCA não ser capaz de gerar estruturas nas quais a representação de relações não taxonômicas é explícita, não significa que tais relações não possam fazer parte dos atributos de contextos formais. Rudolf Wille em [219], por exemplo, gera estruturas FCA a partir de contextos formais em que as relações de incidência (objeto,atributo) são do tipo *(instância,papel_semântico)* e *(instância,classe)*.

3.4 Relational Concept Analysis

Relational Concept Analysis (RCA), proposto por Uta Priss em sua tese de doutorado [168], é uma extensão do método FCA. Por meio do RCA pode-se incluir, na estrutura conceitual, outras relações entre objetos e atributos que não são hierárquicas.

Formalmente, os dados de um RCA são organizados em uma estrutura chamada família de contextos relacionais (FCR). Uma FCR, denotada pelo par (K, R), compreende um conjunto K de contextos formais da forma $K_i = (G_i, M_i, I_i)$ e um conjunto R de relações binárias $r \subseteq G_i \times G_j$ ou $r \subseteq M_i \times M_j$, as quais estabelecem ligações entre os objetos ou atributos dos contextos K_i (domínio) e K_j (imagem) [89]. As relações r entre objetos ou entre atributos são transformadas em relações entre conceitos.

Segundo Priss em [168], ao estender as relações aos conceitos é importante analisar a questão da quantificação, ou seja, verificar se as relações valem para todos os objetos e atributos presentes nas extensões e intensões dos conceitos, ou para apenas um subconjunto deles. Por esta razão, Priss define formalmente a relação entre conceitos incluindo quantificadores.

Desta forma, para um contexto formal (G, M, I), uma relação entre objetos $r \subseteq G \times G$ é transformada em uma relação $\mathcal{R}^r$ entre conceitos $c_1, c_2 \in \mathscr{B}(G, M, I)$, onde $c_1 = (O_1, A_1)$, $c_2 = (O_2, A_2)$ e Q^i representa os quantificadores, para $1 \leq i \leq 4$, conforme as seguintes definições [168]:

$$c_1 \mathcal{R}^r [Q^1, Q^2;] c_2 \Longleftrightarrow Q^1_{g_1 \in O_1} Q^2_{g_2 \in O_2} : g_1 r g_2 \tag{1}$$

$$c_1 \mathcal{R}^r [; Q^3, Q^4] c_2 \Longleftrightarrow Q^3_{g_2 \in O_2} Q^4_{g_1 \in O_1} : g_1 r g_2 \tag{2}$$

$$c_1 \mathcal{R}^r [Q^1, Q^2; Q^3, Q^4] c_2 \Longleftrightarrow c_1 \mathcal{R}^r [Q^1, Q^2;] c_2 \text{ e } c_1 \mathcal{R}^r [; Q^3, Q^4] c_2 \tag{3}$$

Dependendo, portanto, dos quantificadores usados em cada relação r, diferentes relacionamentos $\mathcal{R}^r$ são estabelecidos entre os conceitos. Priss chama r de "componente relacional" e

$[Q^1, Q^2;]$, $[;Q^3, Q^4]$, ou $[Q^1, Q^2; Q^3, Q^4]$ de "etiqueta quantificacional" de uma relação. Para compor as etiquetas quantificacionais, a autora usa, em [168], quantificadores em linguagem natural, tais como $||todos||$,$||ao\,menos\,1||$ e $||exatamente\,1||$. Define, também, abreviações para essas etiquetas, suprimindo, principalmente, quantificadores universais. Por exemplo, as relações $\mathcal{R}^r[||todos||, ||todos||; ||todos||, ||todos||]$ e $\mathcal{R}^r[||todos||, Q^2; ||todos||, Q^4]$ podem ser representadas, respectivamente, por $\mathcal{R}^r_0$ e $\mathcal{R}^r_{[Q^4;Q^2]}$. Para simplificar ainda mais a notação, alguns quantificadores são reescritos em notação matemática. O quantificador $||ao\,menos\,1||$ pode ser representado como $|| \geq 1||$.

A autora ressalta ainda que, se não houver ambiguidades, tanto o componente relacional quanto a etiqueta quantificacional podem ser omitidos na notação da relação [168]. Cabe destacar também que as relações entre atributos $r \subseteq M_i \times M_j$ são tratadas de forma análoga [169].

Para exemplificar as definições apresentadas, vamos representar a relação transversal "é-rival-de", que relaciona um clube ao seu principal concorrente. Nesse caso, a FCR é composta:

- pelo contexto formal derivado $(G, \cup_{m\in M} M_m, J)$, que descreve características de alguns clubes de futebol brasileiros (Tabela 3.4), e
- pela relação $é - rival - de \subseteq G \times G$, mostrada na Figura 3.4a, que relaciona esses clubes aos seus principais concorrentes.

Quanto às etiquetas quantificacionais, para a relação $é - rival - de$, Q^1 e Q^2 poderiam corresponder, respectivamente, aos quantificadores $||todos||$ e $|| \geq 1||$. Desta forma, a relação entre conceitos $c_1 R^{é-rival-de}[||todos||, || \geq 1||;]c_2$, para $c_1, c_2 \in \mathscr{B}(G, \cup_{m\in M} M_m, J)$, definida em (1), estabelece que, para todo clube que pertence à extensão de um conceito c_1, há pelo menos um clube na extensão de c_2 que é seu rival.

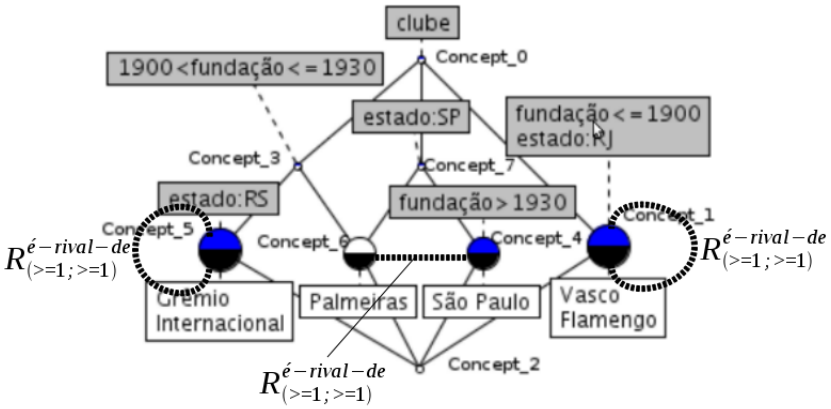
Como a relação $é - rival - de$ é simétrica, no caso da definição em (2), os quantificadores Q^3 e Q^4 poderiam também corresponder, respectivamente, aos quantificadores $||todos||$ e $|| \geq 1||$, ou seja, para todo clube que pertence à extensão de c_2, há pelo menos um clube rival em c_1. Já a equivalência em (3) define apenas a conjunção das duas anteriores.

A Figura 3.4b ilustra[3]a relação $R^{é-rival-de}[||todos||, || \geq 1||; ||todos||, || \geq 1||]$ cuja anotação abreviada é $R^{é-rival-de}_{(\geq 1;\geq 1)}$.

Relação "é-rival-de" entre clubes

	Internacional	Grêmio	São Paulo	Palmeiras	Flamengo	Vasco
Internacional		x				
Grêmio	x					
São Paulo				x		
Palmeiras			x			
Flamengo						x
Vasco					x	

(a)



(b)

Figura 3.4 – Relação "é-rival-de" entre clubes de futebol

Após a formalização da relação, Bendaoud *et al.*, em [15], dão continuidade ao processo de construção da estrutura conceitual, definindo uma "escala relacional". Essa escala tem por

[3]Cabe ressaltar que nesta Figura a relação transversal $R^{é-rival-de}_{(\geq 1;\geq 1)}$ foi editada. A ferramenta Concept Explorer não é apropriada para representar tal relacionamento.

objetivo transformar as relações entre os conceitos $r : K_i \rightarrow K_j$ em atributos de K_i. O resultado dessa transformação é um conjunto de atributos relacionais do tipo $r : c$, onde c é um conceito em K_j. Assim, para um dado objeto $g \in G_i$, o atributo relacional $r : c$ caracteriza a correlação, definida pelos quantificadores, de $r(g)$ com a extensão do conceito c.

Desta forma, para o contexto formal $K = (G, \cup_{m \in M} M_m, J)$ e para a relação $é-rival-de \subseteq G \times G$, definimos a escala relacional $S_{é-rival-de}$ como (G, M^+, J^+), onde:

- $M^+ = \cup_{m \in M} M_m \cup \{é - rival - de : c | c \in \mathscr{B}(K)\}$ e
- $J^+ = J \cup \{(g, é - rival - de : c) | g \in G, c = (O, A) \in \mathscr{B}(K), é - rival - de(g) \cap O \neq \emptyset$.

A partir dessa escala relacional podemos reunir as propriedades locais (Tabela 3.4) e relacionais dos clubes, gerando o contexto formal derivado que é apresentado na Tabela 3.5. No entanto, o uso dessa escala é apenas um passo dentro do processo de análise e construção do reticulado de conceitos. Esse processo é iterativo, visto que o uso de uma escala relacional modifica o contexto formal e o reticulado correspondente, exigindo a aplicação de uma nova escala relacional. Isso se torna necessário, pois novos conceitos são identificados e, portanto, novas relações $r : c$ devem ser inseridas no contexto formal, provocando, assim, atualizações nas intensões dos conceitos já existentes. O processo iterativo cessa quanto um ponto fixo é encontrado, ou seja, a escala relacional aplicada gera um contexto formal cujo reticulado correspondente não se modifica.

Tabela 3.5 – Contexto formal derivado a partir da escala $SR_{é\text{-}rival\text{-}de}$

	clube	fundaçãoAntesDe1901	fundaçãoDe1901A1930	fundaçãoDepoisDe1930	estado:RS	estado:SP	estado:RJ	é-rival-de:Concept_0	é-rival-de:Concept_1	é-rival-de:Concept_2	é-rival-de:Concept_3	é-rival-de:Concept_4	é-rival-de:Concept_5	é-rival-de:Concept_6	é-rival-de:Concept_7
Internacional	x		x		x			x			x		x		
Grêmio	x		x		x			x			x		x		
São Paulo	x			x		x		x			x			x	x
Palmeiras	x		x			x		x				x			x
Flamengo	x	x					x	x	x						
Vasco	x	x					x	x	x						

Esses refinamentos podem ser percebidos analisando-se o reticulado[4] de conceitos da Figura 3.5, que é o resultado desse processo iterativo. Como pode-se observar na Figura 3.4b, existiam inicialmente 8 conceitos, no entanto com a inclusão dos atributos relacionais, surgiu um nono conceito, o *Concept_ 8 = ({São Paulo, Grêmio, Internacional},{clube, é-rival-de:Concept0, é-rival-de:Concept3})*. A identificação desse novo conceito, por sua vez, provocou uma atualização nas intensões do conceito *Concept_ 3 = ({Palmeiras, Grêmio, Internacional},{clube, fundaçãoDe1901A1930, é-rival-de:Concept0, é-rival-de:Concept8})*, que foi a inclusão do atributo relacional *é-rival-de:Concept8.*

Em nosso exemplo, usamos uma FCR simples com apenas um contexto formal e uma relação. No entanto, as FCR podem conter vários contextos e relações. Nesses casos, o resultado

[4] O reticulado de conceitos apresentado nesta figura foi gerado pela ferramenta ERCA. Mais detalhes sobre essa ferramenta podem ser encontrados na Seção 5.3.5. Cabe ressaltar que foi necessário mudar a representação dos atributos de data de fundação, substituindo os operadores relacionais por texto, para que o pacote gráfico usado pela ferramenta gerasse a imagem corretamente. Os rótulos de conceitos *{Concept_ 0, Concept_ 1,...}* são gerados automaticamente pela ferramenta ERCA e, por uma questão de uniformidade, foram usados também nas demais figuras do texto.

do processo é uma família de reticulados relacionais (FRR) cujos conceitos refletem todos os atributos compartilhados e os relacionamentos existentes entre os objetos da FCR [15].

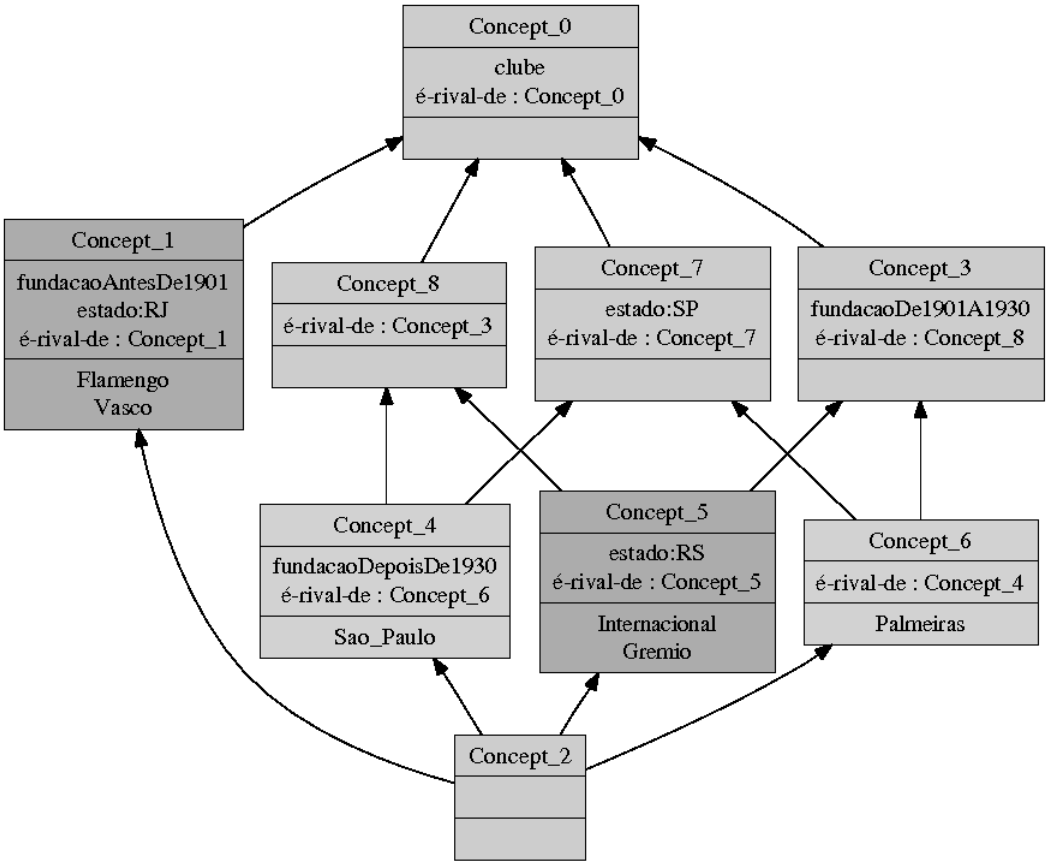


Figura 3.5 – Reticulado de conceitos gerado para o termo "clube" pelo método RCA

Mesmo com FCR mais simples, como a apresentada no exemplo, pode-se perceber que o método produz muitos conceitos. Nas FRR esse problema tende a aumentar. Isso pode se tornar um inconveniente para a estrutura conceitual pretendida, pois tal estrutura pode conter conceitos redundantes ou mesmo ser ilegível (não interpretável) para o usuário, devido ao seu tamanho. Além disso, tende a aumentar também o custo computacional para a geração do reticulado de conceitos. Para minimizar esses problemas, novos algoritmos para geração das estruturas conceituais têm sido propostos. Além disso, é comum o uso de técnicas que permitam diminuir a complexidade de representação dos reticulados, selecionando ou agrupando de alguma forma os seus conceitos.

Na próxima seção comentamos alguns algoritmos para geração de reticulados de conceitos e na seguinte, algumas técnicas usadas para reduzir a complexidade desses reticulados.

3.5 Algoritmos para geração de reticulados de conceitos

Os algoritmos para FCA convivem com dois problemas: a formação do conjunto de conceitos formais e a geração do reticulado de conceitos [117]. Como ambos os problemas demandam processamento com crescimento exponencial, pesquisadores têm estudado a complexidade desses algoritmos [117], bem como proposto diferentes abordagens para a construção de reticulados de conceitos [70, 141, 207].

De acordo com Fu e Nguifo em [70], os algoritmos para construção de reticulados podem ser divididos em algoritmos incrementais e não incrementais.

3.5.1 Algoritmos incrementais

A ideia dos algoritmos incrementais é, a cada novo conceito identificado, atualizar a estrutura. Para isso, os algoritmos fazem, geralmente, uma interseção entre o novo conceito e os conceitos já existentes na estrutura, a fim de determinar onde e como tal conceito deve ser inserido [70].

O algoritmo de Norris foi um dos primeiros algoritmos incrementais para construção de reticulados [70]. Esse algoritmo constrói o reticulado nível a nível. O primeiro nível da estrutura L_1 contém somente o primeiro conceito (O_1, O_1') identificado. Adicionando-se um objeto O_{k+1} ao nível L_k, constrói-se o nível L_{k+1} da seguinte forma [70]:

- $\forall (O_i, O_i') \in L_k$, se $O_i' \subset O_{k+1}'$ então $(O_i \cup O_{k+1}, O_i') \in L_{k+1}$, ou seja, para todos os conceitos i da camada k cujos atributos estiverem contidos nos atributos do novo elemento $k+1$, um novo conceito será gerado com a extensão resultante da união das extensões de i e $k+1$; e com a intensão de i.

- Em caso contrário, $(O_i, O_i') \in L_{k+1}$, e se $(O_i, O_i' \cap O_{k+1}')$ é um maximal[5], esses conceitos devem ser atualizados com as características do conceito $k+1$: $(O_i \cup O_{k+1}, O_i' \cap O_{k+1}')$.

- Depois de examinar todos os conceitos gerados, se O_{k+1}' é um maximal, deve-se adicionar também a L_{k+1} o conceito (O_{k+1}, O_{k+1}').

Godin é outro exemplo de algoritmo incremental. Ele usa uma função *hash*, cuja cardinalidade é definida pelas intensões. Essa função mapeia os conceitos em uma espécie de repositórios, reduzindo assim a pesquisa de conceitos já inseridos na estrutura que tenham algum tipo de similaridade com o novo conceito identificado [117].

Na literatura têm surgido nos últimos anos novos algoritmos incrementais, tais como AddItem em [141] e também versões paralelas que combinam algoritmos incrementais e não incrementais como PSTCL em [95].

3.5.2 Algoritmos não incrementais

Os algoritmos não incrementais (ou *batch*) geram novamente a estrutura a cada novo conceito inserido e podem gerar conceitos basicamente de três formas [70, 117]:

- Geração descendente: os algoritmos partem de um conceito *top* (superconceito) e definem os nodos-filhos (subconceitos) desse conceito. O processo é repetido para os nodos-filhos até gerar a estrutura completamente.

- Geração ascendente: os algoritmos são aglomerativos, ou seja, reúnem conceitos para formar superconceitos.

- Geração por enumeração: os algoritmos enumeram todos os nodos de acordo com algum critério de ordem e usam essa organização para gerar a estrutura.

O algoritmo de Chein é um dos primeiros algoritmos para construção de reticulados [18]. Ele usa a estratégia ascendente *bottom-up*. Os pares objeto-atributo inicialmente identificados (O_i, A_i) são considerados como a primeira camada L_1 do reticulado. Para cada par de elementos (O_i, A_i) e (O_j, A_j) da camada L_k, o algoritmo procede iterativamente da seguinte forma com o objetivo de construir a camada L_{k+1} [70]:

- Se a interseção das intensões não existe ainda na próxima camada $A_i \cap A_j \notin L_{k+1}$, então o algoritmo gera um conceito $(O_i \cup O_j, A_i \cap A_j)$ cuja extensão é a união dos objetos e a intensão, a interseção dos atributos.

[5] A definição de maximal é apresentada no Apêndice A.

- Em caso contrário $A_i \cap A_j \in L_{k+1}$, o algoritmo faz uma junção (*merge*) com os conceitos em L_{k+1} que possuem a mesma intensão $A_i \cap A_j$.

- Ao final de cada iteração k, são removidos os conceitos de L_k cujos atributos são idênticos aos de elementos da camada L_{k+1}.

Bordat e Ganter também são algoritmos não incrementais que surgiram na década de 80. Bordat é um algoritmo *top-down* que consiste basicamente em encontrar os objetos máximos do conjunto de conceitos, ou seja, os seus limites superiores ou o supremo, se ele existir. A partir desses objetos máximos, o algoritmo procura todos os seus subconjuntos. Os subconjuntos se tornam, então, os novos objetos máximos e o processo se repete, construindo desta forma a hierarquia de conceitos [70].

Já o algoritmo de Ganter, chamado também de NextClosure, possui uma abordagem enumerativa. É o mais conhecido, principalmente por ser considerado o mais eficiente dentre os algoritmos mencionados [70, 115, 127]. O algoritmo organiza os conceitos conforme seus atributos, estabelecendo uma ordem lexicográfica que acelera a construção da estrutura conceitual.

Nos últimos anos, têm surgindo extensões desses algoritmos, como ScalingNextClosure [70] e novos algoritmos, como IETreeCS [127]. Têm sido pesquisadas também técnicas de filtragem de conceitos para reduzir a complexidade dos reticulados de conceitos. Essas técnicas são o assunto da próxima seção.

3.6 Técnicas de *Data Weeding*

Conforme Priss e Old, em [173], a geração de reticulados de conceitos grandes e complexos é um desafio para o método FCA. A quantidade de conceitos e de relacionamentos acabam dificultando tanto a exibição quanto a navegação pelos usuários nessas estruturas conceituais. Para contornar esse problema, é comum o uso de técnicas de filtragem, denominadas "data weeding". Essas técnicas reduzem a complexidade das estruturas conceituais, considerando necessidades específicas e o propósito da aplicação. Priss e Old organizam essas técnicas em 4 grupos: técnicas de redução visual; técnicas utilizam *faceting* e *plaine scales*; técnicas de poda e restrição; e técnicas de decomposição e que usam *general scales*.

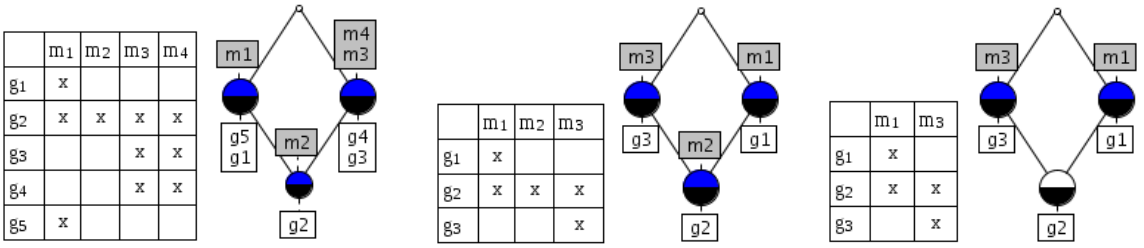
3.6.1 Redução visual

As técnicas de redução visual não modificam o reticulado de conceitos gerado pelo método FCA, apenas alteram a sua forma de exibição. Logo, não há perda de informação. A maioria das técnicas desse grupo permitem a omissão de objetos e de atributos. O critério que determina a omissão desses elementos varia conforme a técnica em uso [173].

Na técnica de *clarification,* descrita por Ganter em [75], por exemplo, os objetos que têm exatamente os mesmos atributos podem ser substituídos por um único objeto. O mesmo pode ser feito com os atributos que possuem a mesma extensão.

Já na técnica de redução, também descrita por Ganter em [75], são omitidos os atributos que são equivalentes a combinações de outros atributos. Um atributo $m \in M$ de um contexto em que a técnica de *clarification* foi aplicada é redutível, se há outros atributos $m_1, m_2 \in M$, diferentes de m, tais que a interseção de suas extensões resulta na extensão de m ($m'_1 \cap m'_2 = m'$). Caso contrário, ele é dito irredutível. A técnica também pode ser usada para reduzir os objetos do contexto.

A Figura 3.6 ilustra as técnicas descritas. O contexto formal e o correspondente reticulado apresentados na Figura 3.6a são transformados nas estruturas da Figura 3.6b, após a técnica de *clarification*. E estes, por sua vez, são transformados na Figura 3.6c, após a aplicação da



(a) (b) (c)

Figura 3.6 – Redução visual do reticulado obtida a partir das técnicas de *clarification* e redução.

redução. É importante notar que a estrutura do reticulado não é modificada, apenas os rótulos dos nodos.

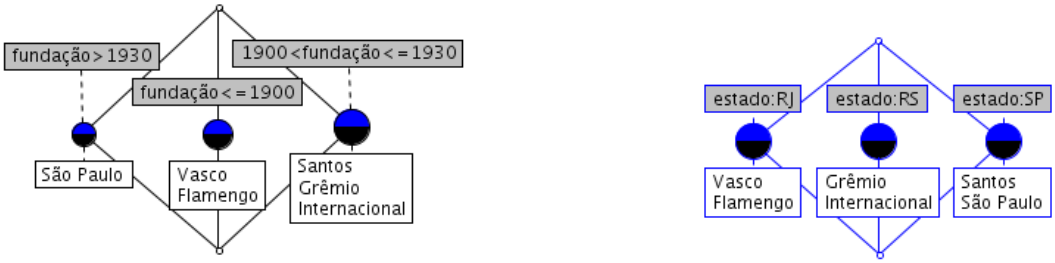
Estão incluídas também neste grupo de redução visual as técnicas de *zoom* e aquelas que permitem deslocar graficamente, de alguma forma, tanto a estrutura quanto seus nodos.

3.6.2 *Faceting* e *plaine scale*

Faceting e plaine scale permitem subdividir o reticulado de conceitos em reticulados menores [173]. Segundo Priss, em [170], o termo *faceting* é aplicado na área de ciência da informação para designar o processo de estruturação de conceitos de forma hierárquica. Por exemplo, um livro pode ser classificado como "Ficção/ReinoUnido/Século19", usando as características (*facets*): tópico, local e tempo. A autora trata os termos *faceting* e *plaine scale* como equivalentes quando usadas para FCA, visto que as estruturas conceituais resultantes da aplicação de tais técnicas são muito similares [170]. Como Priss considera o termos equivalentes, nos deteremos apenas em *plaine scale*.

As *plaine scales* foram apresentados na Seção 3.3 quando descrevemos contextos formais multivalorados. Como elas não promovem a perda de informação, desde que a divisão dos dados seja criteriosa, também podem ser usadas para reduzir a complexidade das estruturas. Assim, para cada subdivisão pretendida, define-se uma escala correspondente. Cada escala permitirá construir apenas uma parte do reticulado original.

Para exemplificar, subdividimos o reticulado para clubes de futebol apresentado na Figura 3.3 em dois, usando as escalas $S_{dataDeFundação}$ (Figura 3.7a) e S_{estado} (Figura 3.7b), ambas definidas na Seção 3.3.



(a) (b)

Figura 3.7 – Reticulados de conceitos gerados a partir das escalas $S_{dataDeFundação}$ e S_{estado}.

A Figura 3.8 ilustra a representação do reticulado de conceitos para clubes após a aplicação dessas escalas.

3.6.3 Poda e restrição

Técnicas de poda são aquelas que removem objetos, atributos ou conceitos da estrutura conceitual. Em muitos casos, as mudanças realizadas por tais técnicas são tão significativas que

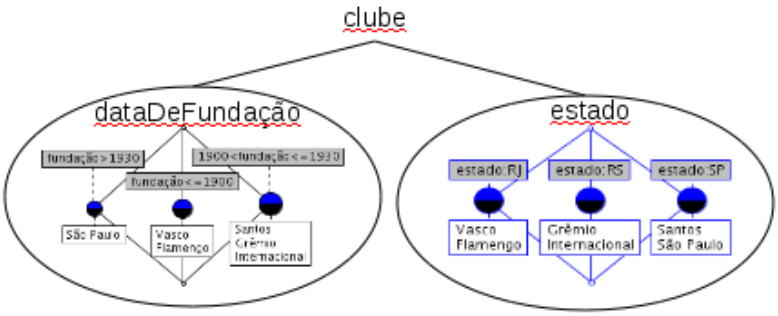


Figura 3.8 – Reticulado de conceitos gerado após a aplicação da técnica de *plaine scale*

a estrutura conceitual resultante deixa de ser um reticulado, tornando-se apenas um conjunto ordenado [173].

A seleção de conceitos promovida pela poda é efetuada a partir de algum índice conhecido. Todos os conceitos cujos índices não atingirem um valor limite pré-definido pelo usuário são descartados. Um dos índices de poda mais citados baseia-se na noção de estabilidade [57, 112, 117, 173].

A estabilidade de um conceito pode ser medida tanto de forma intensional quanto extensional. O índice de estabilidade intensional σ_i, definido na Equação 3.1, estabelece o quanto a intensão de um conceito depende de determinados objetos da extensão. A idéia é determinar quantos objetos são necessários e suficientes para criar um conceito [112]. O índice de estabilidade intensional σ_e, definido na Equação 3.2, dualmente estabelece o quanto a extensão de um conceito depende de seus atributos.

$$\sigma_i(A,B) = \frac{|\{C \subseteq A | C' = B\}|}{2^A} \tag{3.1}$$

$$\sigma_e(A,B) = \frac{|\{D \subseteq B | D' = A\}|}{2^B} \tag{3.2}$$

Para exemplificar, calculamos para os conceitos do reticulado da Figura 3.9b os índices de estabilidade σ_i e σ_e. Para isso, tivemos que reverter a etiquetagem reduzida aplicada ao reticulado. A Figura 3.9a apresenta a tabela com valores calculados para os índices de estabilidade, conforme o algoritmo apresentado no Anexo A. Se o limite definido pelo usuário fosse $0,7$, por exemplo, apenas os conceitos c_3 e c_4 seriam selecionados como estáveis. Há autores, como Falk *et al.* em [57], que eliminam também os conjuntos que possuem menos de dois objetos como extensão ou menos de dois atributos como intensão. Nesse caso, apenas o conceito c_3 seria selecionado.

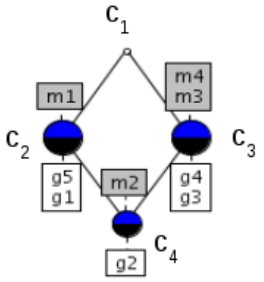
Diferente das técnicas de poda, as técnicas de restrição não exigem a construção completa do reticulado de conceitos para serem aplicadas. Elas são usadas durante a construção dos chamados "reticulados de vizinhança" (*neighbourhood lattices*). Um reticulado de vizinhança é um subconjunto do reticulado original. Seu processo de construção consiste basicamente em escolher um objeto (ou atributo) e buscar itens relacionados a ele. Um operador, chamado "operador mais", é usado para repetir esse processo um número determinado de vezes. Assim, a cada aplicação do "operador mais", novos itens relacionados aos já existentes são recuperados [172].

A técnica de restrição é usada para modificar o "operador mais", estabelecendo a quantidade (dois, três, etc) de atributos (ou objetos) que os novos objetos (ou atributos) recuperados devem ter em comum com os já existentes no reticulado, para serem selecionados [173].

Estabilidade dos Conceitos

	conceito	σ_i	σ_e
c_1	$\{g_1, g_2, g_3, g_4, g_5\}, \{\}$	0.56	1
c_2	$\{g_1, g_2, g_5\}, \{m_1\}$	0.75	0.5
c_3	$\{g_2, g_3, g_4\}, \{m_3, m_4\}$	0.75	0.75
c_4	$\{g_2\}, \{m_1, m_2, m_3, m_4\}$	1	0.75

(a)



(b)

Figura 3.9 – Exemplo de cálculo de estabilidade de conceitos

3.6.4 Decomposição e *general scale*

As técnicas de decomposição particionam o reticulado de conceitos em reticulados menores [173]. A técnica de decomposição horizontal (Figura 3.10), por exemplo, consiste em excluir os elementos $\perp$ e $\top$ do reticulado de conceitos, gerando, assim, subreticulados disjuntos $L_1, L_2, ..., L_n$, onde $\perp \leq x$ e $\top \geq x$ para todo $x \in \sum_{i=1}^{n} L_i$. Para que essa técnica possa ser aplicada, a soma horizontal desses subreticulados, definida em 3.3, deve resultar no reticulado original [71].

$$\sum_{i=1}^{n} L_i = \{\top, \perp\} \cup U_{i=1}^{n} L_i \setminus \{\top_i, \perp_i\} \tag{3.3}$$

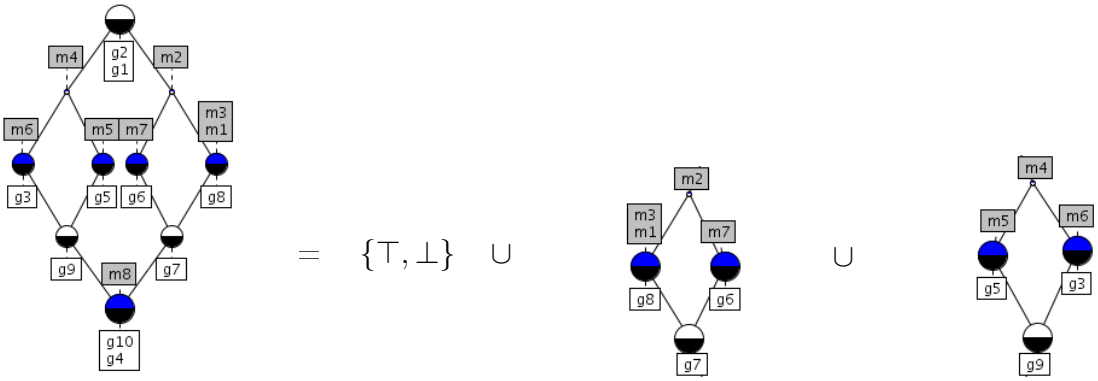


Figura 3.10 – Exemplo de decomposição horizontal

Já as *general scales* transformam os dados do reticulado em "dados mais genéricos". Utilizam, para isso, algum "operador de composição" que mapeie o conjunto original de termos em termos mais abrangentes. Priss e Old comentam em [173] que o uso de tais escalas não é uma prática muito comum. Por isso não lhes será dado maior destaque aqui.

Embora não tenham sido citadas pelos autores, as medidas de similaridade aplicadas a termos ou conceitos também podem colaborar para reduzir a complexidade do reticulado. Em aplicações de construção de estruturas conceituais a partir de textos, tais medidas são mais utilizadas. No entanto, não propriamente para reduzir a complexidade do reticulado, mas para minimizar o problema da esparsidade das informações obtidas a partir de textos (esse problema foi comentado na Seção 2.5.2). Como essas informações não são completas, visto que os textos não possuem todas as dependências possíveis entre os termos, as medidas permitem a união de termos considerados semelhantes, constituindo assim conceitos mais densos e reticulados mais compactos.

A próxima seção apresenta algumas medidas de similaridade usadas em aplicações de construção de estruturas conceituais a partir de textos.

3.7 Similaridade entre termos e conceitos

A esparsidade da informação para aplicações que utilizam textos como fonte de extração dessas informações não é uma novidade. Manning e Schütze em [135] comentam esse problema e propõem uma forma de atenuá-lo por meio de um método ao qual chamaram *smoothing*. O método *smoothing* consiste, essencialmente, em atribuir probabilidades diferentes de zero para eventos não observados nos textos.

Cimiano *et al.* em [41] baseiam-se nesse método para definir a "similaridade mútua". Dois termos t_1 e t_2 são considerados mutuamente similares se, conforme uma medida de similaridade previamente estabelecida, t_1 for o termo mais similar a t_2 e, reciprocamente, t_2 for o termo mais similar a t_1. Os autores usaram medidas bem conhecidas, como cosseno e Jaccard [123], para definir a similaridade entre os objetos (nomes) a partir de seus atributos (verbos). Os objetos considerados mutuamente similares foram agrupados e tiveram seus atributos compartilhados. Desta forma, relações entre objetos e atributos que não estavam presentes nos textos passaram a existir no contexto formal da estrutura FCA. Os autores comentam ainda que, embora o cosseno tenha produzido índices de similaridade mais adequados, a técnica "similaridade mútua" de maneira geral não produziu bons resultados.

De acordo com Formica em [63], o coeficiente Dice tem sido usado na literatura para comparar conjuntos de atributos e, desta forma, estabelecer a similaridade entre os objetos. Esse é o caso do trabalho de Otero *et al.* em [157] que usam o coeficiente Dice para comparar atributos formados por contextos lexicossintáticos.

Embora a similaridade entre conceitos ontológicos seja amplamente estudada, Formica também afirma que há poucos trabalhos que pesquisem a similaridade entre os conceitos formais de um FCA [63]. Alqadash e Bhatnagar em [4] ressaltam que, enquanto os conceitos em ontologias são expressos como "rótulos" de dados, os conceitos formais não têm rótulos. Eles contêm simplesmente conjuntos de objetos (as extensões dos conceitos) e atributos (os descritores desses conceitos). Esta diferença influencia nos resultados e, por isso, justifica a necessidade de medidas de similaridade específicas para FCA.

Nas duas próximas subseções apresentamos duas medidas de similaridade recentemente desenvolvidas para conceitos formais.

3.7.1 Índice zeros-induced

Alqadash e Bhatnagar em [4] propõem um índice de similaridade para conceitos formais denominado zeros-induced. Diferente de medidas mais usuais, este índice considera a informação compartilhada pelos conceitos. Para os conceitos formais $C_1 = (G_1, M_1)$ e $C_2 = (G_2, M_2)$ obtidos a partir do contexto formal $\mathscr{B}(G, M, I)$, o índice zeros-induced é calculado como definido na Equação 3.4. A função $z(C_1, C_2)$, que aparece na definição, corresponde ao número de zeros inclusos (atributos ausentes) na submatriz de $\mathscr{B}(G, M, I)$ induzida a partir das linhas $|G_1 \cup G_2|$ e colunas $|M_1 \cup M_2|$.

$$S_z(C_1, C_2) = \frac{|G_1 \cup G_2| * |M_1 \cup M_2| - z(C_1, C_2)}{|G_1 \cup G_2| * |M_1 \cup M_2|}, \tag{3.4}$$

$$\text{onde } z(C_1, C_2) = \textstyle\sum_{g \in G_1 \cup G_2} |(M_1 \cup M_2) \diagdown g')|$$

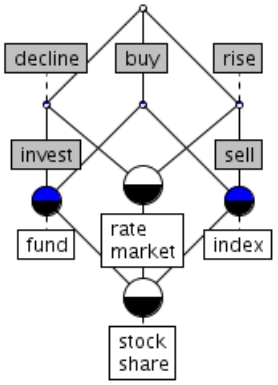
Para exemplificar, considere que o objetivo é calcular o índice zeros-induced para os pares de conceitos $S_z(C_1, C_2)$ e $S_z(C_1, C_3)$, onde $C_1 = (\{share, stock, index\}, \{buy, sell, rise\})$, $C_2 = (\{share, stock, fund\}, \{buy, decline, invest\})$ e $C_3 =$ *({share, stock, market, rate}, {rise, decline})*. Esses conceitos pertencem ao reticulado da Figura 3.11b e foram obtidos a partir do

contexto formal da Figura 3.11a. As informações de dependências entre verbos e argumentos usadas nesse exemplo foram extraídas do *corpus* PropBank (Seções 4.3 e 5.1.1.2).

Contexto Formal

	buy	*sell*	*rise*	*decline*	*invest*
share	1	1	1	1	1
stock	1	1	1	1	1
market	0	0	1	1	0
fund	1	0	0	1	1
rate	0	0	1	1	0
index	1	1	1	0	0

(a)



(b)

Figura 3.11 – Contexto formal e reticulado gerado a partir do PropBank.

Na submatriz gerada para calcular $z(C_1, C_2)$, apresentada na Figura 3.12a, há 4 atributos ausentes, logo $S_z(C_1, C_2) = 16/20 = 0,8$. Já na submatriz $z(C_1, C_3)$, mostrada na Figura 3.12b, há 8 atributos ausentes, logo $S_z(C_1, C_3) = 17/25 = 0,68$. Portanto, o índice zeros-induced determinou que os conceitos C_1 e C_2 são mais similares. Se estivéssemos usando uma medida que utilizasse apenas a cardinalidade dos atributos em comum, C_1 seria considerado semelhante a C_2 e a C_3 na mesma proporção.

Submatriz de $z(C_1, C_2)$

	buy	*sell*	*rise*	*decline*	*invest*
share	1	1	1	1	1
stock	1	1	1	1	1
fund	1	0	0	1	1
index	1	1	1	0	0

(a)

Submatriz de $z(C_1, C_3)$

	buy	*sell*	*rise*	*decline*	*invest*
share	1	1	1	1	1
stock	1	1	1	1	1
market	0	0	1	1	0
rate	0	0	1	1	0
index	1	1	1	0	0

(b)

Figura 3.12 – Exemplo de submatrizes geradas para o cálculo do índice zeros-induced.

Alqadash e Bhatnagar em [4] compararam o índice a três outras medidas, dentre elas a medida Jaccard, para diferentes bases de dados. Em todos os experimentos em que os dados eram esparsos, o índice zeros-induced obteve melhores resultados. A desvantagem do índice, no entanto, é que seu custo computacional é quadrático $O(n^2)$ enquanto que o das demais medidas é linear $O(n)$.

3.7.2 Medida Sim

Formica em [63] propõe uma medida para comparar conceitos formais cuja base é a medida de similaridade ics (*information content similarity*). A medida ics foi introduzida por Resnik em [179] e refinada por Lin em [128]. De acordo com essa medida, quanto mais informações os conceitos compartilharem, mais similares eles serão.

A medida ics, cuja definição é apresentada em 3.5, combina abordagens tradicionais baseadas na distância entre os conceitos em uma taxonomia tal como a WordNet (contribuição de Resnik) com a probabilidade de tais conceitos aparecerem em um *corpus* (contribuição de Lin) [128].

$$ics(c_1, c_2) = \frac{2 * log(p(c'))}{log(p(c_1)) + log(p(c_2))}, \text{ onde } c' = lso(c_1, c_2) \text{ e } p(n) = \frac{freq(n)}{M} \tag{3.5}$$

A função $lso(c_1, c_2)$ determina na taxonomia o "menor" ancestral que os conceitos c_1 e c_2 têm em comum. Já $p(n)$ tem como propósito determinar a probabilidade de um conceito n aparecer em um *corpus*. Essa probabilidade corresponde ao quociente da frequência absoluta de n pelo número total M de instâncias de conceitos que existem no *corpus*.

A partir da medida ics Formica, em [63], propõe a medida Sim, definida em 3.6. Ela calcula a similaridade entre os conceitos formais $C_1 = (G_1, M_1)$ e $C_2 = (G_2, M_2)$ de um mesmo contexto ou de diferentes contextos formais.

$$Sim(C_1, C_2) = \frac{|(G_1 \cap G_2)|}{g} \times w + \frac{s(M_1, M_2)}{m} \times (1 - w) \tag{3.6}$$

Em (9), g e m correspondem às maiores cardinalidades dos conjuntos G_1 e G_2, M_1 e M_2, respectivamente. O valor do peso w é definido pelo usuário e, segundo Formica, foi introduzido para flexibilizar o método quanto à importância de objetos e atributos no cálculo da similaridade. A função s utiliza a medida ics para calcular a similaridade máxima dos atributos. Para determinar o valor de s, é necessário inicialmente gerar um conjunto P, que é um subconjunto de $M_1 \times M_2$. O conjunto P deve conter os pares de atributos mais similares conforme a medida ics, sendo que não devem ser incluídos nesse conjunto pares que compartilhem atributos. Assim, para cada atributo em M_1 deve-se encontrar o mais similar em M_2.

Para exemplificar, considere que o objetivo é calcular a medida Sim para os mesmos pares de conceitos apresentados na seção anterior. Vamos considerar, neste exemplo, o peso w como $0,5$, valorizando assim, tanto os objetos quanto os atributos. Para calcular $Sim(C_1, C_2)$, onde $C_1 = (\{share, stock, index\}, \{buy, sell, rise\})$, $C_2 = (\{share, stock, fund\}, \{buy, decline, invest\})$, começamos definindo o conjunto P de pares, tal que a soma da medida[6] ics dos seus elementos seja máxima. O conjunto P, nesse caso, é $\{(buy, buy), (sell, invest), (rise, decline)\}$, pois atinge a similaridade máxima de $1,79$, que é o valor de s (Tabela B.1 do Apêndice B). Assim, a similaridade entre os conceitos C_1 e C_2 é definida como

$$Sim(C_1, C_2) = \frac{2}{3} * 0,5 + \frac{1,79}{3} * 0,5 = 0,63$$

Para calcular a $Sim(C_1, C_3)$, onde $C_3 = (\{share, stock, market, rate\}, \{rise, decline\})$, o conjunto P que maximiza s é $\{(buy, decline), (rise, rise)\}$, onde s é $1,06$ (Tabela B.2 do Apêndice B). Nesse caso, a similaridade entre os conceitos C_1 e C_3 é calculada como

$$Sim(C_1, C_3) = \frac{2}{4} * 0,5 + \frac{1,06}{3} * 0,5 = 0,42$$

Como podemos observar, a medida Sim também considerou os conceitos C_1 e C_2 como mais similares e apresenta um custo computacional igualmente quadrático $O(n^2)$. No entanto, segundo Formica em [63], a medida Sim apresenta vantagens. Ela pode produzir melhores resultados que o coeficiente Dice, principalmente por usar a medida ics que, de acordo com o mesmo autor, provê avaliações de similaridade próximas ao julgamento humano.

Dado que a complexidade do reticulado de conceitos é um problema, e que medidas para reduzir sua complexidade e calcular a similaridade dos conceitos são necessárias, comentamos na seção seguinte vantagens e desvantages de se utilizar o FCA como método de agrupamento conceitual.

[6] O cálculo da medida *ics* foi realizado com o pacote NLTK (Seção 5.3.2), usando-se a WordNet 3.0 como taxonomia e o Brown como *corpus* de referência. Para mais informações, consulte o Apêndice B.

3.8 FCA como método de agrupamento conceitual

De acordo com Cimiano *et al.* em [40], a escolha de um método de agrupamento para construção de taxonomias a partir de textos deve-se basear nos seguintes aspectos: eficácia (qualidade do resultado), eficiência (comportamento em tempo de execução) e rastreabilidade da construção da estrutura conceitual pelo engenheiro. Na visão de Cimiano *et al.,* o método FCA atende a todos esses aspectos.

A principal vantagem do método FCA em comparação a métodos de agrupamento baseados em similaridade é que, além de gerar os grupos de conceitos, ele também provê uma descrição intensional para esses grupos. Conforme Cimiano *et al.* [40], essa descrição intensional facilita a rastreabilidade do processo de construção da estrutura ontológica e, de acordo com Zhang *et al.*, tal descrição torna os grupos gerados melhor interpretáveis [232].

Já a sua principal desvantagem é a complexidade do reticulado de conceitos gerado pelo método [40, 232]. No pior caso, o desempenho do método FCA é exponencial $O(2^n)$, enquanto que os métodos baseados em similaridade são geralmente quadráticos $O(n^2)$ [40]. Na prática, no entanto, tal complexidade não é atingida para aplicações que buscam informações a partir de textos, em razão do problema da esparsidade das informações. Cimiano *et al.* em [40], por exemplo, avaliaram o método FCA, em seus experimentos para construção de estruturas conceituais a partir de textos, com desempenho próximo ao linear. Zhang *et al.* em [232] contornaram o problema da complexidade usando heurísticas para selecionar os conceitos mais relevantes.

Valtchev e Missaoui em [209], por outro lado, acreditam que o FCA e os métodos de agrupamentos baseados em similaridade são complementares. Eles compararam os métodos a partir de 6 critérios: o tipo de similaridade usada para definir os conceitos, a singularidade da saída gerada, a forma de agrupamento (particionamento ou de sobreposição), a especialização entre conceitos, o uso de pesos nos atributos e a capacidade de manipular diferentes tipos de dados.

Diferente Cimiano *et al.,* Valtchev e Missaoui consideram a noção de similaridade melhor definida em métodos de agrupamentos baseados em similaridade, por usarem medidas mais precisas. Cimiano *et al.* usam justamente esse argumento em favor do método FCA. Segundo Cimiano *et al.,* os atributos são capazes de justificar de forma mais compreensível a formação dos grupos (conceitos) do que mera medidas numéricas.

De acordo com Valtchev e Missaoui, o método FCA tem uma uniformidade no seu processo, ou seja: a sua forma de agrupamento, que define a generalização/especialização dos conceitos, é baseada na inclusão dos objetos (relação subconceito-superconceito definida na Seção 3.2). Desta forma, o método produz, para os mesmos dados, sempre uma saída singular. Já nos demais métodos de agrupamento, esses aspectos, especialmente a saída, dependerão do algoritmo e da configuração usada para esse algoritmo (parâmetros).

Valtchev e Missaoui criticam o método FCA ainda por não permitir a atribuição de pesos aos atributos que formam os conceitos. Outra crítica refere-se à dificuldade de se trabalhar com dados numéricos. Para representá-los na estrutura conceitual é necessária primeiramente a definição de escalas conceituais (Seção 3.3).

Apesar das desvantagens do método, muitos autores consideram que suas vantagens ainda são maiores e justificam o uso do FCA e de suas extensões em aplicações voltadas para a construção de estruturas conceituais a partir de textos. Algumas dessas aplicações são apresentadas na próxima seção.

3.9 Aplicações do método na geração de estruturas conceituais a partir de textos

Cimiano *et al.* em [41] propõem uma abordagem baseada no método FCA para construir ontologias a partir de textos dos domínios Turismo e Finanças. Para cada domínio, os autores

extraem de textos, em inglês e em alemão, relações de dependência entre os verbos e seus argumentos. Para encontrar essas relações de dependência, os autores marcam os textos com o lematizador e etiquetador de POS TreeTagger (comentado na Seção 5.3.1) e o *parser* LoPar[7]. As relações identificadas são lematizadas e filtradas. Diferentes medidas de informação, em [39], são testadas pelos autores para selecionar as dependências mais relevantes, sendo que a probabilidade condicional é mencionada como a medida que obteve os melhores resultados.

Após esse processo de seleção, os autores aplicam a técnica de "similaridade mútua" (apresentada na Seção 3.7) para determinar os termos mais similares e agrupá-los. As relações resultantes são então transformadas em um contexto formal a partir do qual o reticulado de conceitos é gerado. O reticulado ainda passa por uma etapa de poda, na qual são removidos os nodos internos cuja extensão é a mesma de um nó filho. O objetivo da poda é remover conceitos mais abstratos, deixando assim a estrutura mais compacta. O resultado da poda é uma ordem parcial que se aproxima muito de uma hieraquia de conceitos.

As hierarquias geradas para os domínios Turismo e Finanças são comparadas a ontologias construídas manualmente por meio das medidas Taxonomic Ovelap e semantic cotopy, comentadas na Seção 2.6.2.2. Os índices produzidos pela avaliação foram baixos. Para o domínio Turismo, a precisão foi 29,33% e o *recall*, 65,49%. E para Finanças, a precisão e o *recall* foram, respectivamente, 29,93% e 37,05%.

Otero *et al.* em [157] também usam o FCA para gerar estruturas conceituais a partir de textos. Os autores realizam seus experimentos com *corpora* em português e em inglês, os quais, logo de início, são marcados, respectivamente, pelos etiquetadores TreeTagger e Freeling[8] com informações de POS. Dos textos etiquetados, são extraídos nomes (substantivos) e os contextos lexicossintáticos desses nomes.

Um contexto lexicossintático é definido como um padrão linguístico constituído por uma palavra, uma relação sintática e uma posição morfossintática. O padrão "president of [NOUN]" é um exemplo de contexto lexicossintático para nomes que denotam entidades com um presidente, tais como "Portugal", "Belgium", "Real Madrid", "company" e "republic". Segundo os autores, o uso de contextos lexicossintáticos é mais eficiente do que palavras, pois são menos polissêmicos. Os autores descartam os contextos lexicossintáticos esparsos e agrupam os mais semelhantes utilizando o coeficiente Dice como medida de similaridade.

Para aplicar o método FCA, os autores definem duas operações: especificação e abstração. As operações são aplicadas a pares de objetos e atributos, representados por (O, A), onde os objetos são os nomes e os atributos, os contextos lexicossintáticos. Na operação de especificação, são gerados pares mais restritos. Nessa operação, os pares (O_1, A_1) e (O_2, A_2), cujos contextos lexicossintáticos são similares, produzem o par $(O_1 \cap O_2, A_1 \cup A_2)$. Já na operação de abstração, são gerados pares mais gerais, da forma $(O_1 \cup O_2, A_1 \cap A_2)$. A estrutura conceitual resultante é analisada por avaliadores humanos, segundo critérios estabelecidos pelos próprios autores. A acurácia com que os nomes foram associados aos contextos lexicossintáticos ficou em torno de 92% [157] .

Jia *et al.* em [98] utilizam o FCA para construir ontologias usando 900 artigos científicos sobre armazenamento e recuperação de informações da biblioteca digital da ACM. Diferente de Cimiano *et al.*, eles extraem dados de categorização que aparecem em artigos ACM, formando pares do tipo *(palavras-chave, classificação)* para definir os contextos formais e, conseqüentemente, caracterizar os conceitos. Os autores descrevem também como mapear o reticulado para ontologias em RDF. E ainda apresentam uma aplicação em que a estrutura é usada para a expansão de consultas. Segundo os autores, o método permitiu enriquecer as consultas com palavras-chave mais relevantes.

Bendaoud *et al.* em [16] utilizam um tesauro (taxonomia NCBI - National Center for

[7] http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar.html
[8] http://gramatica.usc.es/pln/tools/freeling.html

Biotechnology Information), uma base de dados (sobre bactérias[9]), um *corpus* (1.244 artigos da PubMed[10]) para gerar ontologias na área de microbiologia a partir dos métodos FCA e RCA.

Inicialmente, os autores constróem reticulados conceituais FCA de domínio. Para construir o reticulado sobre bactérias, por exemplo, os autores extraem do tesauro as classes dessas bactérias, e da base de dados e do *corpus*, pares do tipo *(objeto, atributo)*, onde os objetos correspondem às bactérias e os atributos, as suas propriedades. Em seguida, eles geram dois reticulados conceituais: um relacionando as bactérias às suas classes e outro relacionando as bactérias às suas propriedades. Por fim, os reticulados são unidos, formando um só. O mesmo processo é feito para gerar o reticulado conceitual sobre antibióticos. O papel do RCA é relacionar os reticulados de domínio, identificando, por exemplo, relações de resistência de certas bactérias a alguns antibióticos.

A avaliação deste esforço, descrita em um trabalho posterior [17], segue a metodologia de Cimiano e coautores em [41] e, portanto, faz uso de uma ontologia de referência. Embora os resultados sejam ainda preliminares, os autores perceberam que a precisão do método FCA foi maior que a do RCA.

No trabalho de Hacene *et al.* em [89], o método RCA é usado no processo de construção de uma ontologia para o domínio de Astronomia. Inicialmente, os autores usam o Stanford *parser*[11] para marcar 830 *abstracts* de artigos do jornal A&A obtidos na base de dados SIMBAD [12]. São extraídas, dos textos etiquetados, as relações entre os verbos e seus argumentos. São consideradas apenas as relações em que os nomes que aparecem em tais argumentos estão definidos em um tesauro de Astronomia[13]. Os autores utilizam ferramenta GATE para refinar as relações extraídas e escolher as mais relevantes considerando as suas frequências. Para determinar a similaridade entre os nomes compostos encontrados no texto, os autores usam uma adaptação da medida de similaridade definida por Wu e Palmer em [224]. Os valores de similaridade dos termos são então avaliados por um especialista no domínio.

A partir das relações resultantes desse processo, Hacene *et al.* constróem reticulados de conceitos com o método FCA. Eles formam dois contextos formais. Um para os corpos celestiais identificados, relacionando os nomes desses corpos celestiais (objetos formais) aos verbos que representam as suas propriedades (atributos formais). E outro contexto para os telescópios, relacionando os nomes desses telescópios às suas características (perigeu, período orbital e massa). A seguir, os autores criam uma família de contextos relacionais, que estabelecem relacionamentos transversais entre os corpos celestiais e os telescópios que podem observá-los, segundo a tecnologia utilizada por esses telescópios. A partir da família de contextos relacionais é gerado então o RCA.

Os autores definem ainda um conjunto de regras de tradução para representar os elementos do RCA em lógica de descrição e, portanto, gerar uma ontologia. A ontologia resultante foi avaliada de forma empírica por avaliadores humanos peritos em Astronomia. Segundos os avaliadores, a maioria das classes encontradas correspondem de fato a categorias de corpos celestiais conhecidas, no entanto os relacionamentos transversais entre os telescópios e corpos celestiais nem sempre foram julgados significativos. Hacene *et al.* consideram os resultados ainda preliminares e atribuem a má qualidade de certos relacionamentos à ausência de elementos, visto que, em seus experimentos, foram usados apenas 10 diferentes tipos de telescópios e 60 diferentes corpos celestiais.

Analisando-se os trabalhos pesquisados, podemos perceber que os autores costumam utilizar informações lexicossintáticas para gerar os contextos formais. Geralmente, são extraídas dos

[9] http://bac.hs.med.kyoto-u.ac/jp
[10] http://ncbi.nlm.nih.gov/sites/entrez
[11] http://nlp.stanford.edu/software/lex-parser.shtml. A seção 5.3.3 descreve a versão mais atual desse *parser*.
[12] http://simbad.u-strasbg.fr/simbad/
[13] http://msowww.anu.edu.au/library/thesaurus/

textos as relações entre os verbos e seus argumentos, e nem sempre a ambiguidade dos termos é tratada. Cimiano *et al.* em [41], por exemplo, calcula a frequência dos pares (nome,verbo) sem considerar que nomes e verbos podem ser polissêmicos. O trabalho de Otero *et al.* [157] é um dos poucos que explicitam a preocupação com a polissemia usando contextos lexicossintáticos ao invés de palavras.

O processo de seleção dos padrões mais relevantes é baseado na frequência desses padrões no texto. No entanto, não há uniformidade quanto ao uso dessa frequência para medir a importância de cada padrão. Não há consenso também quanto aos valores mínimos que esses padrões devem atingir, para serem selecionados.

Embora os autores citem o coeficiente Dice como a medida mais usual para determinar a similaridade entre os conceitos [63, 157], entre os trabalhos estudados percebemos que medidas de similaridade que utilizam a WordNet também são utilizadas [89].

O uso de ontologias de referência no processo de avaliação é recorrente, provavelmente por permitir a comparação entre diferentes métodos de construção de estruturas conceituais de forma automática. No entanto, é comum o uso de juízes humanos na análise qualitativa dos conceitos. Isso se deve, principalmente, à dificuldade de estabelecer comparações entre os elementos extraídos de um *corpus* e os conceitos existentes em uma ontologia de referência. Até porque, como já foi mencionado, os conceitos formais são constituídos de extensões e intensões, e os conceitos ontológicos, como vistos na ciência da computação, são "rótulos" que generalizam um conjunto de dados.

Embora os resultados de alguns trabalhos com RCA ainda sejam preliminares, aparentemente o método FCA como método de agrupamento para gerar estruturas conceituais a partir de textos tem provido resultados mais satisfatórios que o RCA.

3.10 Considerações sobre este capítulo

O uso do método FCA na construção de estruturas ontológicas e em aplicações na área de sistemas de informação tem crescido nos últimos anos. Encontramos trabalhos recentes em conferências sobre ontologias, como Formal Ontologies in Information Systems (FOIS), e em periódicos de inteligência artificial, como Journal of Artificial Intelligence Research (JAIR). Mas é importante destacar que conferências internacionais dedicadas ao FCA existem desde a década de 90. As principais conferências relacionadas o método são International Conference on Conceptual Structures (ICCS), Concept Lattices and Their Applications (CLA) e International Conference on Formal Concept Analysis (ICFCA).

Em ciência da computação, o interesse por FCA e suas extensões se justifica visto que é um método adequado para a análise de dados e tem se mostrado promissor para a representação de conhecimento. Sua inerente característica de relacionar e organizar os dados de forma hierárquica [231, 167] o torna um método de agrupamento conceitual muito interessante para construção de estruturas ontológicas. Isso ocorre especialmente por prover descrições intensionais dos dados que facilitam a interpretação dos agrupamentos gerados [98] e por facilitar a identificação de relações não taxonômicas. Sua potencialidade está em permitir diferentes representações conceituais que refletem as diversas formas como os dados aparecem relacionados nos textos. Sendo indicado, portanto, para análises linguísticas, pois gera estruturas que possibilitam o estudo de relacionamentos sintáticos e semânticos, inclusive para desambiguação de sentido [167].

Por outro lado, é importante o uso de boas heurísticas para seleção dos pares *(objeto, atributo)*, pois, como a estrutura gerada é na verdade um grafo, as possibilidades de combinação dos elementos cresce exponencialmente, à medida que o volume de pares aumenta. Embora grafos representem melhor o complexo relacionamento das entidades no mundo [98], visto que não trabalhamos apenas com relações de ordem puramente taxonômica, a complexidade dos

algoritmos é um fator a ser considerado. Um outro aspecto pouco mencionado, mas importante quanto à representação, é que o método não trata a negação. Podendo ser, assim, inadequado para aplicações em que essa situação deve ser representada.

Embora os resultados obtidos pelos pesquisadores com o método sejam interessantes e motivadores, ainda não é de nosso conhecimento a existência de trabalhos, na atualidade, que usem o método FCA para geração de estruturas conceituais a partir de textos e que incluam aspectos semânticos, mais especificamente classes de verbos e papéis semânticos, nos seus contextos formais. Para que pudéssemos construir a nossa proposta combinando o método FCA com tais aspectos, fizemos um estudo sobre papéis semânticos o qual é apresentado no próximo capítulo.

4. PAPÉIS SEMÂNTICOS

Este capítulo introduz papéis semânticos, apresentando seu conceito e os tipos de etiquetas semânticas geralmente usadas pelas ferramentas de anotação. Comenta trabalhos relacionados à organização dos verbos em classes semânticas e também faz uma breve apresentação dos principais recursos lexicais envolvidos em tarefas de anotação, assim como de alguns etiquetadores automáticos de papéis semânticos. São apresentadas ainda, algumas aplicações de papéis semânticos na construção de estruturas conceituais a partir de textos.

4.1 Conceito e tipos de anotações semânticas

Papéis semânticos, também chamados papéis temáticos ou ainda de papéis-θ[1] (θ-*roles*), tipicamente expressam a relação semântica entre um predicado[2] e seus argumentos [120, 161]. Tais papéis foram introduzidos na década de 60 através de trabalhos como o de Fillmore [61], sob a justificativa que relações sintáticas eram insuficientes para representar as relações de dependência existentes entre os participantes de um evento descrito por um verbo.

Os papéis semânticos permitem caracterizar esses participantes quanto às suas ações e estados nos eventos [33]. Essas relações de dependência nas estruturas predicado-argumento podem ser facilmente percebidas analisando-se as sentenças abaixo:

(a) [$John_{Agent}$] *broke* [*the window*$_{Patient}$].

(b) [*The window*$_{Patient}$] *broke*.

(c) [John$_{Agent}$] opened [the door$_{Patient}$].

(d) [The key$_{Instrument}$] opened [the door$_{Patient}$].

Semanticamente, tanto na sentença (a) quanto na (b) o sintagma *the window* (a janela) é quem sofre a ação definida pelo verbo *to break* (quebrar), desempenhando, desta forma, o papel de Patient (paciente) do verbo *to break*. O efeito do evento *to break* provocará uma mudança de estado na janela, que presumidamente ao final desse evento ficará danificada. Além disso, diferente do que acontece na sentença (b), em (a) a entidade que efetuou a ação está claramente definida, ou seja, John é o Agent (agente) de *to break*.

Cabe ressaltar que, embora (a) e (b) estejam em voz ativa, *the window* possui funções sintáticas distintas nessas sentenças: em (a) ele é o objeto direto de *to break* e em (b), o seu sujeito. O verbo *to break* apresenta também um comportamento sintático diferenciado nas sentenças: em (a) está funcionando como transitivo e em (b) como intransitivo. Já no caso das sentenças (c) e (d) sintaticamente não há diferença, John e *key* (chave) são sujeitos, no entanto em (c) John é o agente do verbo *to open* (abrir) e em (d) *key* é o seu instrumento (Instrument foi o papel associado ao objeto utilizado na ação descrita pelo verbo). Como podemos observar, a análise sintática não é o bastante para explicitar o significado das sentenças [159].

[1] Embora alguns autores usem o termo papel-θ para indicar relações de caráter puramente sintático [33], neste documento seguiremos autores como Jackendoff [96] que trata os termos papel-θ, papéis temáticos e papéis semânticos como sinônimos.

[2] Papéis semânticos podem ser atribuídos a argumentos de predicados nominais [161], no entanto nos deteremos apenas em predicados verbais por serem os mais mencionados pelos pesquisadores e, principalmente, pelo fato de os etiquetadores automáticos, em geral, anotarem apenas os argumentos de verbos.

Os papéis Agent, Patient e Instrument apresentados nas sentenças fazem parte de um conjunto de papéis para os quais existe uma certa concordância entre os linguístas quanto às suas características. No entanto não há uma lista consensual de papéis semânticos. Há linguístas, inclusive, que questionam se de fato essa lista poderia existir [33, 139].

As listas propostas geralmente se referem a papéis para situações específicas ou a um conjunto pequeno de papéis mais gerais. Normalmente, quando os papéis possuem rótulos mais específicos, portanto mais incomuns e menos consensuais, os autores costumam chamá-los simplesmente de papéis semânticos [159]. Já no caso dos papéis mais gerais e mais conhecidos, o termo mais usado é mesmo papéis temáticos [97].

Alguns dos papéis temáticos mais comuns na literatura, são abordados a seguir (as sentenças apresentadas como exemplos foram extraídas da página *web* do pesquisador Sowa[3]).

- **Agent** (agente): é associado frequentemente ao sujeito da sentença [61, 97, 159]; corresponde a uma entidade, tipicamente humana ou pelo menos animada, que provoca uma ação ou evento [61] de forma voluntária [99, 162]. Alguns autores também atribuem o papel de agente a objetos (máquinas, por exemplo) [110] e a forças da natureza (chuva, vento, ...) [97, 99]. Exemplo: *[Eve $_{Agent}$] bit an apple.*

- **Patient** (paciente): entidade diretamente afetada por uma ação, mudando o seu estado [96]. Pode ser animado ou inanimado. Sintaticamente costuma ocorrer como objeto dos verbos [97, 110]. Exemplo: *[The cat $_{Agent}$] swallowed [the canary $_{Patient}$].*

- **Instrument** (instrumento): objeto geralmente inanimado que participa de forma secundária da ação, sendo também uma causa do evento descrito pelo verbo [99, 120]. Pode ser usado pelo agente. Tipicamente não muda o seu estado [199] mas pode provocar mudança de estado de outras entidades após o evento [110]. É frequentemente introduzido pela preposição *with* (com). Exemplo: *[The key $_{Instrument}$] opened the door.*

- **Theme** (tema): refere-se à entidade que muda de posição (local) [97]. Em geral, aparece associado ao objeto do verbo *to give* (dar) e aos sujeitos dos verbos *to walk* (caminhar) e *to die* (morrer) [110]. Alguns autores, no entanto, usam esse papel para indicar uma entidade que funciona como uma espécie de gatilho a partir do qual uma outra entidade passa a experimentar um estado. Neste caso, pode aparecer relacionado a verbos que expressam sensações e sentimentos [33, 161]. Jurafsky e Martin descrevem esse papel como o participante mais afetado por um evento [99]. Exemplo: *Billy likes [the Beer $_{Theme}$].*

- **Source** (fonte): é o elemento a partir do qual a ação se inicia [96, 120]. Jurafsky e Martin atribuem a mesma semântica a esse papel, o qual definem como a origem de um movimento [99]. Exemplo: *The chapter begins [on page 20 $_{Source}$].*

- **Destination** (meta) **/Recipient** (recipiente): Destination é o oposto de Source, refere-se ao elemento para o qual o movimento se dirige [96, 120]. Aparece geralmente como sujeito dos verbos *to receive* (receber) e *to buy* (comprar) [110]. É usado também para indicar uma mudança de posse (propriedade). Alguns autores, entretanto, costumam utilizar o papel de Recipient para este fim, ou seja, o papel é usado em entidades concretas ou abstratas que são alvos de uma transferência [110]. Exemplos : *Bob went [to Danbury $_{Destination}$]. Sue sent the gift [to Bob $_{Recipient}$].*

- **Experiencer** (experienciador): pessoa afetada por um estado ou evento (sensorial, cognitivo ou emocional); é entendido como um tipo de paciente [96]. Vários verbos de caráter emocional - *to love* (amar), *to admire* (admirar), ... - e psicológico - *to amuse* (divertir),

[3] http://www.jfsowa.com/ontology/thematic.htm

to perturb (pertubar), ... - costumam ter seus argumentos (sujeito ou objeto) associados a esse papel [110]. Exemplo: *[Yojo $_{Experiencer}$] sees the fish.*

- **Beneficiary** (beneficiário): usado para indicar a entidade que se beneficia com uma ação [99, 120], geralmente introduzido pela preposição *for* (para) [110]. Alguns autores não fazem distinção entre os papéis beneficiário e recipiente (ou meta) [156]. Exemplo: *Diamonds were given [to Ruby $_{Beneficiary}$].*

- **Location** (local): corresponde à entidade (lugar) em que o evento acontece [61, 120]; é geralmente introduzido por sintagmas preposicionais [110]. Exemplo: *Vehicles arrive [at a station $_{Location}$].*

Mesmo no caso de papéis mais gerais, como os temáticos, para os quais se percebe um certo consenso, há algumas divergências quanto à quantidade e à caracterização. Dowty [51] tenta minimizar essas divergências propondo, então, apenas dois papéis semânticos, aos quais chamou de **proto-agente** e **proto-paciente**. O **proto-agente** tem propriedades que abrangem vários dos papéis citados e se caracteriza por: participar de forma voluntária no evento, causar um evento ou mudança de estado em outros elementos participantes do evento, mover-se, existir independentemente do evento e outras. Já o **proto-paciente** tem a propriedade de: suportar mudanças de estado, ser afetado por outro elemento participante do evento, não existir independentemente do evento, etc.

Em abordagens mais atuais esse impasse quanto à definição dos papéis tem sido contornada, de uma certa forma, pela atribuição de rótulos numéricos ($Arg0$, $Arg1$, $Arg2$,...) aos argumentos dos verbos [159]. Esse é o caso do PropBank (Seção 4.3 e 5.1.1.2), um *corpus* anotado manualmente que tem sido muito utilizado principalmente para treinar etiquetadores automáticos de papéis semânticos.

O problema é que a anotação provida pelo *corpus* e, consequentemente, por muitos etiquetadores semânticos não é uniforme quando os verbos são de classes diferentes. Apesar de existir uma determinada regularidade para os rótulos $Arg0$ e $Arg1$, que costumam corresponder, respectivamente, aos papéis agente e paciente, o mesmo não se pode afirmar para os demais rótulos. Na anotação PropBank, o rótulo $Arg2$, por exemplo, para os verbos *to kick* (com sentido de chutar) e *to slice* (com sentido de fatiar) corresponde, respectivamente, aos papéis instrumento e fonte [159]. Como a classe dos verbos é importante para esse tipo de anotação, esta é o assunto da próxima seção.

4.2 Classes de verbos

O comportamento homogêneo dos verbos quanto à flexão e função sintática permite estabelecer, sob o ponto de vista morfossintático, classificações bem definidas para os verbos. No entanto, o caráter polissêmico e a variedade de construções em que os verbos podem aparecer nas sentenças dificultam generalizações semânticas e, consequentemente, classificações desta ordem.

Um dos trabalhos mais mencionados na literatura quanto à classificação semântica dos verbos é o de Beth Levin [125]. A autora correlaciona os verbos semanticamente a partir de seus comportamentos sintáticos [94]. Apesar da relação "comportamento sintático" e "significado" não ser perfeita, existem regularidades suficientes para a formação de classes [139, 214]. Verbos que permitem as mesmas alternativas de construção de sentenças em nível sintático (transitivo, intransitivo, ...) compartilham características semânticas em algum sentido e, portanto, podem ser agrupados [126].

As relações entre os verbos de uma classe, no entanto, não são necessariamente de sinonímia [126]. Algumas classes, como *Break* (*to break* - quebrar, *to chip* - lascar, *to crack* - rachar, *to fracture* - fraturar, *to rip* - rasgar, ...), contêm verbos cujo sentido é muito próximo, mas em

outras, como *Braid* (*to braid* - trançar, *to brush* - escovar, *to clip* - prender, *to comb* - pentear, *to curl* - enrolar...), nem tanto [110]. Por este motivo, alguns autores consideram essa classificação difusa, já que não estabelece claramente o tipo de relação semântica entre os verbos [10, 184].

Por outro lado, os verbos de uma mesma classe possuem estruturas de argumentos similares com uma certa coerência semântica. Por exemplo, verbos que denotam eventos, como *to cut* (cortar), *to kill* (matar) e *to destroy* (destruir), são geralmente transitivos e possuem uma entidade agente, funcionando como sujeito, que atua e causa a mudança em uma entidade paciente, que é um objeto direto [126].

Essas e outras regularidades encontradas por Levin têm servido como base para o desenvolvimento de ferramentas automáticas e semiautomáticas para classificação de verbos [214], etiquetagem de papéis semânticos [82], desambiguação de sentido dos verbos [47], etc. E também como ponto de partida para criação de recursos lexicais, como PropBank [159] e VerbNet [110].

Os recursos frequentemente mencionados na literatura para anotação de papéis semânticos incluem o *corpus* Proposition Bank (PropBank) e os léxicos FrameNet e WordNet. O léxico VerbNet também é usado com esse fim em alguns trabalhos. Por esta razão, esses recursos léxicos são brevemente comentados nas seções a seguir.

4.3 PropBank

O Proposition Bank, ou simplesmente PropBank, é um *corpus* anotado com papéis semânticos muito mencionado, especialmente, em trabalhos referentes a etiquetadores semânticos. De acordo com Palmer, Kingsbury e Gildea em [159], o PropBank foi criado a partir do Treebank-2 (Seção 5.1.1). O TreeBank-2 corresponde a aproxidamente um terço do Penn TreeBank. Ele é formado essencialmente por textos em língua inglesa do Wall Street Journal. Sendo, portanto, do domínio de Finanças.

O PropBank, que foi etiquetado manualmente, provê anotações semânticas às estruturas predicado-argumento [109]. O verbo de uma sentença geralmente indica um evento particular e os participantes desse evento estão associados aos argumentos sintáticos desse verbo [108]. As anotações semânticas propostas pelos autores são inspiradas em trabalhos como o de Levin (Seção 4.2) e o de Dowty [51].

Em razão da dificuldade de definir um conjunto universal de papéis semânticos, o PropBank define um conjunto de papéis, chamado *roleset,* para cada uso distinto de um verbo. Os papéis atribuídos aos argumentos dos verbos são rotulados por números de 0 (zero) a 5 (cinco): $Arg0$, $Arg1$, $Arg2$,...

Os autores usam também um rótulo especial chamado $ArgA$ para agentes cuja ação é induzida, ou seja, para os casos em que a ação executada pelo agente é motivada por uma força externa. É o caso, por exemplo, de verbos como *to march* (marchar).

Além dos *rolesets*, há rótulos para indicar as funções dos argumentos nas sentenças. Tipicamente esses argumentos são atribuídos a elementos adjuntos $ArgMs$ como por exemplo, o modo de um verbo (MOD). A Tabela 4.1 descreve os rótulos $ArgMs$ definidos pelos autores.

Aos *rolesets* são associadas as estruturas sintáticas compostas pelos argumentos permitidos para cada verbo, bem como alguns exemplos (sentenças anotadas), formando um *frameset* [159]. A Figura 4.1 mostra o *frameset* 01 do verbo *to accept* (aceitar) no sentido *to take willingly* (receber espontaneamente). A estrutura sintática definida para esse uso do verbo é caracterizada pelos argumentos: $Arg0$ indicando quem aceitou (*Acceptor*), $Arg1$ determinando o que foi aceito (*Thing accepted*), $Arg2$ definindo a origem do que foi aceito (*Accepted-from*) e $Arg3$ descrevendo os atributos do que foi aceito (*Attribute*).

Tabela 4.1 – Subtipos de rótulos *ArgMs* [159].

Rótulo	Descrição	Rótulo	Descrição
LOC	local	CAU	causa
EXT	grandeza	TMP	tempo
DIS	conectivo de discurso	PNC	propósito
ADV	propósito geral	MNR	maneira
NEG	marcador de negação	DIR	direção
MOD	modo do verbo		

Frameset accept.01 "*take willingly*"
*Arg*0: *Acceptor*
*Arg*1: *Thing accepted*
*Arg*2: *Accepted-from*
*Arg*3: *Attribute*
Ex : ... [$_{Arg0}$*He*] [$_{ArgM-MOD}$*would*] [$_{ArgM-NEG}$*n't*] *accept* [$_{Arg1}$*anything of value*] [$_{Arg2}$*from those he was writing about*].

Figura 4.1 – *Frameset accept.01* do PropBank [159].

Já no caso de verbos polissêmicos em que o uso pode estabelecer diferenças de significado, há geralmente mais de um *frameset*. Um dos fatores que define um novo *frameset*, por exemplo, é a necessidade de associar mais argumentos ao verbo em um determinado uso. Ao conjunto de *framesets* de um verbo, os autores chamam de *frames file* (arquivos de *frames*). Os *frames files* formam o léxico de verbos do PropBank. A Figura 4.2 mostra os *framesets* 01 e 02 do verbo *to decline*, respectivamente no sentido *to go down incrementally* (em queda) e no sentido *to demure, to reject* (rejeitar).

Frameset *decline.01* "go down incrementally"
*Arg*1: *entity going down*
*Arg*2: *amount gone down by*, EXT
*Arg*3: *start point*
*Arg*4: *end point*
Ex : ... [$_{Arg1}$*its net income*] *declining* [$_{Arg2-EXT}$*42%*] [$_{Arg4}$*to $121 milion*] [$_{ArgM-TMP}$*in the first 9 month of 1989*].

Frameset decline.02 "*demure, reject*"
*Arg*0: *agent*
*Arg*1: *reject thing*
Ex : ... [$_{Arg0}$*A spokesman*] *declined* [$_{Arg1}$*to elaborate*].

Figura 4.2 – *Frameset decline.01* e *decline.02* do PropBank [159].

Embora os autores tenham procurado atribuir o mesmo conjunto de papéis semânticos a verbos que pertençam à mesma classe, ou seja, que compartilham as mesmas estruturas sintáticas e sentido, não há um padrão definido. No entanto, os rótulos *Arg*0 e *Arg*1, na maioria dos casos, correspondem, respectivamente, aos papéis agente e paciente (ou tema) [159].

Inicialmente, dos cerca de 3.300 verbos diferentes existentes no Treebank-2, o PropBank possuía apenas 1.826 deles [159]. Atualmente, o PropBank provê anotação para todos os verbos do Treebank-2, sendo considerado, por esta razão, uma amostra representativa quanto à anotação de papéis semânticos. Sua limitação, no entanto, está no fato de se referir apenas

a um gênero de texto, impossibilitando o desenvolvimento de bons anotadores semânticos para outros domínios [139].

O *corpus* PropBank é distribuído comercialmente pela LDC (Linguistic Data Consortium). Além do léxico de verbos, a LDC fornece a anotação propriamente dita, que corresponde às sentenças TreeBank-2 etiquetadas com os papéis semânticos. O léxico de verbos, embora esteja no pacote da LDC, também pode ser encontrado na página do projeto PropBank[4]. Na Seção 5.1.1.2 há mais detalhes sobre a anotação das sentenças.

4.4 FrameNet

O FrameNet é uma base de dados composta por *frames* semânticos para Língua Inglesa, onde cada *frame* é uma estrutura conceitual que captura a semântica de uma determinada situação. Os *frames,* além de descreverem os itens lexicais (em geral, nomes, adjetivos e verbos) e os papéis semânticos desses itens em uma dada situação, provêem também sentenças anotadas que exemplificam tais descrições [9]. Pelo fato de conter sentenças, o FrameNet é considerado tanto um léxico computacional quanto um *corpus* anotado com papéis semânticos. Ele possui cerca de 141.000 sentenças anotadas manualmente, que foram obtidas principalmente do BNC (British National Corpus) [139].

Os papéis semânticos, no FrameNet, são chamados *frame elements* (FEs) e são locais aos *frames*. Eles descrevem os agentes e os objetos envolvidos na situação semântica caracterizada pelo *frame*. As situações descritas pelos *frames* pertencem a domínios semânticos como corpo (partes e funções), cognição, comunicação, percepção, transações, tempo, espaço e outros [9].

A Figura 4.3 apresenta o frame *Building* (Construção). Este *frame* foi extraído diretamente da página *web* do projeto FrameNet[5]. Ele exemplifica de forma simplificada a estrutura de um *frame*. Um *frame* contém basicamente um identificador (*Building*); uma definição textual (*definition*); FEs principais (*core*) e secundários (*non-core*) juntamente com seus tipos semânticos (*semantic type*) e unidades lexicais (*lexical units*).

Building
Definition*: This frame describes assembly or construction actions, where an Agent joins Components together to form Created_entity, ...* Ex: [$Jack_{Agt}$] *built* [*a new house*$_{CrEnt}$] *with* [*hammer and nails*$_{Ins}$].
FEs:
Core:
Agent [Agt]: builds a Created_entity.
Semantic Type: Sentient
Created_entity[CrEnt]: identifies the entity that is created in the act of building.
Semantic Type: Artifact
...
Non-Core:
Instrument [Ins]: This FE identifies the Instruments(s) with which an Agent builds a Created_entity.
Semantic Type: Physical_entity
...
Lexical Units
assemble.v, assembly.n, build.v, ...

Figura 4.3 – *Frame Building* (extraído da página *web* do projeto FrameNet).

Embora seja considerado um importante recurso linguístico, um dos problemas do FrameNet, que justifica inclusive o fato de não ser este amplamente usado principalmente por etiquetadores automáticos, é que o seu *corpus* não é uma amostra representativa da linguagem, é mais um conjunto de sentenças escolhidas manualmente para ilustrar as possibilidades de

[4] http://www.cis.upenn.edu/~ace
[5] http://framenet.icsi.berkeley.edu

atribuição de papéis semânticos a itens. Outro problema é o uso de papéis mais específicos pode dificultar a implementação dos anotadores automáticos [139].

4.5 VerbNet

VerbNet[6] é um léxico de verbos em inglês. Ele organiza os verbos em classes hierárquicas. Essa organização é definida a partir das propriedades sintáticas e semânticas dos verbos, sendo considerada uma extensão das classes de verbos de Levin (Seção 4.2) [107].

No primeiro nível da estrutura hierárquica estão as classes definidas por Levin que, de acordo com Kingsbury e Kipper em [107], sofreram algumas modificações para garantir a sua uniformidade quanto ao compartilhamento de papéis semânticos. Nos demais níveis da estrutura, as classes vão sendo especializadas. Elas herdam todas as características de suas classes-pai, mas adicionam informações que definem restrições de papéis semânticos e que expandem as estruturas sintáticas e predicados semânticos herdados.

A Figura 4.4 apresenta um exemplo de classe VerbNet a partir da classe Hit-18.1. Cada classe descreve os papéis semânticos e suas restrições. Dos 25 papéis[7] semânticos tratados pelo léxico (ator, agente, beneficiário, causa, local, etc [110]), os argumentos dos verbos da classe Hit-18.1 podem assumir apenas 3: agente, paciente e instrumento.

Os papéis possuem restrições que permitem caracterizar os argumentos quanto ao tipo (concreto ou abstrato), tempo, estado, localização, escala, etc. No caso da classe Hit-18.1, os argumentos definidos como instrumento (*instrument [+concret]*) devem ser concretos [110].

A classe ainda indica o conjunto de verbos que compartilham características. No exemplo, pertencem à classe os verbos *to bang* (bater, atirar), *to bash* (derrubar), *to hit* (acertar, atingir, golpear), *to kick* (chutar) e outros. Os *frames* dessa classe descrevem as características sintáticas e semânticas desses verbos, indicando regência verbal (transitivo, intransitivo, ...), provendo exemplos de sentenças, estabelecendo restrições sintáticas aos papéis semânticos (pontuação, plural, ...) e definindo predicados semânticos quanto à causa, ao movimento, a aspectos temporais, etc [110].

Os argumentos desses predicados semânticos podem ser eventos, constantes, papéis temáticos e verbos específicos. Na classe Hit-18.1, o predicado *cause* indica que o agente do verbo é a causa de um evento identificado como E.

Class Hit-18.1	
Roles and Restrictions: Agent[+int_ control] Patient [+concrete] instrument [+concrete]	
Members: bang, bash, hit, kick, ...	
Frames	
Name	*Basic Transitive*
Example	*Paula hit the ball*
Syntax	*Agent* $\vee$ *Patient*
Semantics	*cause(Agent,E), manner (during(E),directedmotion,Agent), ...*

Figura 4.4 – Classe Hit-18.1 do léxico VerbNet (adaptado de [110]).

O VerbNet contém 5.200 verbos organizados em 237 classes [110]. Esse léxico tem sido considerado um recurso importante na área da linguística computacional. No entanto, para a tarefa de etiquetagem de papéis semânticos, seu uso tem sido mais restrito, principalmente

[6] http://verbs.colorado.edu/~mpalmer/projects/verbnet

[7] Alguns desses papéis estão descritos na Tabela C.4 do Apêndice C.

pela falta de uma associação direta com um *corpus* semanticamente anotado. Embora o VerbNet mapeie alguns verbos do PropBank, para a tarefa de etiquetagem através de abordagens supervisionadas, esse relacionamento não tem sido considerado representativo [139].

4.6 WordNet

A WordNet[8] é uma base lexical que provê o significado de mais de 120.000 palavras, incluindo substantivos, verbos, adjetivos e advérbios em inglês [143]. É uma rede semântica cujos nodos são conjuntos de sinônimos, chamados de *synsets*. Os *synsets* contêm todas as formas (palavras) em que se pode expressar um determinado conceito. Eles contêm também uma glosa de definição e, em alguns casos, exemplos de sentenças sobre esses conceitos. As palavras são ligadas dentro dos *synsets* através de relações sinonímia e antonímia. No entanto, os *synsets* por meio das relações de hiponímia e hiperonímia formam uma estrutura hierárquica de conceitos. Os *synsets* também possuem relações de meronímia e holonímia.

No caso dos verbos, mais especificamente, há relações de hierarquia como troponímia, que conecta os verbos quanto à maneira como eles realizam uma ação. Por exemplo, *to swipe* (bater ou golpear com violência) e *to smack* (dar uma palmada ou tapa) são tropônimos de *to hit* (bater). A relação nesse caso refere-se aos graus de força aplicados na ação. Já em verbos que denotam eventos a relação é mais parecida com hiponímia, conectando conceitos mais gerais a específicos, como no caso de *to plumment* (cair verticalmente) que é tropônimo de *to drop* (cair) [59]. A WordNet ainda contém relações de implicação *(entailment)* e também de causa para verbos [143].

Basicamente, ela é composta de verbos que denotam ações e eventos em 14 domínios semânticos específicos, que denotam movimento, percepção, contato, comunicação e outros. Descreve também verbos que indicam estados, incluindo auxiliares e de controle. Ela provê diferentes características aos verbos. Nos que indicam movimento, por exemplo, ela especifica velocidade, como *to walk* - *to run* (caminhar-correr); direção, como *to rise* - *to fall* (levantar-abaixar); e meios de deslocamento, como *to walk* - *to drive* (caminhar-dirigir) [59].

É um recurso muito utilizado na construção de estruturas ontológicas como mencionado na Seção 2.5.2, e vem sendo usada também na tarefa de anotação de papéis semânticos [202], classificação semântica de verbos [36, 114] e desambiguação semântica de verbos [227].

4.7 Etiquetadores de papéis semânticos

A tarefa de atribuir papéis semânticos automaticamente aos argumentos de um verbo tem sido alvo de pesquisa intensa nos últimos anos. Sendo considerada pelos pesquisadores como uma tarefa importante no processo de compreensão da linguagem [156]. Para esse tipo de anotação, que não é trivial, várias abordagens têm sido propostas.

Abordagens baseadas em modelos estatísticos e em algoritmos de aprendizagem de máquina parecem ser as mais comuns. A maioria dos trabalhos utiliza o PropBank [164] ou mesmo as sentenças do FrameNet [77] como *corpus* de treino. Essas estratégias envolvem heurísticas para atribuir os papéis apoiadas, muitas vezes, em algum recurso léxico (WordNet, FrameNet e VerbNet) ou na combinação deles [78].

Um dos primeiros trabalhos na área, muito citado, é o de Gildea e Jurasfsky [77]. Os autores extraem de 50.000 sentenças do *corpus* FrameNet diferentes características léxicas, sintáticas e posicionais dos constituintes dessas sentenças. Tais características, combinadas com informações probabilísticas quanto à atribuição de papéis, formam a base dos classificadores estatísticos propostos pelos autores para realizar a tarefa de anotação semântica.

[8] http://wordnet.princeton.edu/

Já Swier e Stevenson em [202] apresentam um método não supervisionado para atribuir papéis semânticos. A abordagem dos autores baseia-se em um algoritmo de *bootstrapping* cujo processo de atribuição de papéis inicialmente não considera qualquer ambiguidade na tarefa. Para isso, usa padrões preliminares de papéis definidos a partir das informações semânticas de verbos da VerbNet. Ao longo de sua execução, no entanto, um modelo probabilístico de atribuição de papéis é criado e iterativamente aperfeiçoado à medida que novas anotações são realizadas. O modelo leva em consideração informações sobre a classe dos verbos, bem como sobre as funções sintáticas e a classe dos nomes dos argumentos. A classe dos nomes é definida a partir da WordNet.

Em trabalhos como o de Toutanova *et al.* em [206], que utiliza o PropBank como *corpus* de treino, outras características têm sido incorporadas aos etiquetadores, principalmente no que se refere à relação de dependência entre as etiquetas semânticas de certos argumentos e as funções sintáticas de outros argumentos na sentença. Sendo uma tendência, ao que parece, o uso de modelos mais globais quanto à atribuição de papéis. Para o *corpus* Wall Street Journal, os autores relatam que o modelo proposto obteve na tarefa de anotação 81.9% de precisão, 78.81% de abrangência e um F1 de 80,32%. Já para o *corpus* Brown, o F1 caiu para 68,81%.

Pradman e Ward em [164] descrevem o etiquetador ASSERT que também usa o PropBank como *corpus* de treino. O trabalho dos autores tem uma relação estreita com o de Gildea e Jurasfky [77] quanto às etapas envolvidas no processo de atribuição dos papéis e às características linguísticas usadas para anotação. Inicialmente, o sistema identifica os argumentos dos verbos das sentenças e define um conjunto das possíveis etiquetas semânticas para cada argumento. Por meio de classificadores SVM, são atribuídas aos argumentos as etiquetas apropriadas. Como a etiquetagem é feita de forma local para cada argumento, ao final o sistema combina as etapas anteriores provendo assim os rótulos finais aos argumentos. Os autores conseguem uma pequena melhora no F1 do *corpus* Brown (~70%) ao incluírem de forma incremental pequenas partes desse *corpus* no conjunto de treino.

Para a Língua Portuguesa, o único trabalho que encontramos sobre anotação automática de papéis é o de Bick em [19]. O autor utiliza 500 regras definidas manualmente para anotar sentenças pré-processadas pelo seu *parser* PALAVRAS. A abordagem do autor trabalha com 35 papéis semânticos e foi aplicada sobre parte do *corpus* Floresta Sintá(c)tica Treebank [183]. Para este *corpus*, o autor relata um F1 de 88.6%.

Os etiquetadores de papéis semânticos que utilizamos em nossos estudos são comentados na Seção 5.3.4.

Na seção seguinte, apresentamos algumas aplicações de verbos e papéis semânticos na área de aprendizagem de ontologias a partir de textos.

4.8 Aplicações de verbos e papéis semânticos na aprendizagem e enriquecimento de estruturas ontológicas

Encontramos com frequência trabalhos que não usam FCA mas exploram verbos e papéis semânticos em diferentes tarefas relacionadas à aprendizagem e ao enriquecimento de estruturas conceituais a partir de textos, especialmente no que se refere à extração e identificação de relações transversais entre conceitos.

Esse é o caso do trabalho de Maedche e Staab em [131]. Os autores identificaram os conceitos e seus relacionamentos a partir de dependências sintáticas entre os verbos e seus argumentos em textos *web* do domínio Turismo. Usaram também regras de associação para definir as relações mais relevantes, bem como estabelecer o nível de abstração mais adequado para tais relações. Os autores definiram uma nova medida, a Relation Learning Accuracy ($\overline{RLA}$), para avaliar a similaridade entre as relações não taxonômicas identificadas automaticamente e aquelas elaboradas manualmente. Usaram também medidas de precisão e abrangência para

avaliar a abordagem adotada. O melhor $\overline{RLA}$ obtido foi de 0,67, para um suporte de 0,04 e uma confiança de 0,01 (parâmetros das regras de associação), o que resultou em 13% e 11% de abrangência e precisão, respectivamente. No entanto, a medida $\overline{RLA}$ de 0,53, com suporte de 0,06 e confiança de 0,4 levou a abrangência e a precisão para 0%. Por esta razão, os autores consideraram o método de avaliação proposto fraco, para abordagens automáticas.

Kavalec e Svátek em [104] utilizam uma abordagem semelhante, no entanto propõem a medida Above Expectation (AE), baseada em probabilidade condicional, para estimar a associação entre verbos e pares de conceitos (identificados sintaticamente como argumentos desses verbos). Os autores usaram em seus experimentos os *corpora* em língua inglesa Lonely Planet[9], do domíno de Turismo, e SemCor[10], que é uma parte do *corpus* Brown[11] semanticamente anotado com sentidos da WordNet. Utilizaram ontologias de referência para avaliar seus experimentos. No entanto, para minimizar o problema referente à falta de relações não taxonômicas nessas ontologias, o que acaba penalizando a precisão, os autores usaram duas medidas para este fim: precisão anterior e precisão posterior. A precisão anterior é a tradicionalmente usada. Ela compara as relações descobertas com as existentes na ontologia de referência. Já a posterior é calculada sobre as relações obtidas a partir da abordagem, que não existiam na ontologia de referência, mas que foram consideradas corretas pelos especialistas humanos. Foram criadas também medidas "anterior" e "posterior" para abrangência e *F-measure*. Os melhores resultados para o *corpus* Lonely Planet, em termos de precisão, foi para $AE = 3$. Nesse caso, as precisões anterior e posterior atingiram 100%, embora a abrangência anterior tenha ficado por volta dos 35%. Já para o SemCor, a precisão posterior atingiu 100% quando foi usado $AE = 4$.

Sanchez e Moreno em [187] também centralizam nos verbos o processo de descoberta de relações transversais a partir de textos *web*. A metodologia prevê tanto a aprendizagem de relações taxonômicas quanto não taxonômicas. A aprendizagem de relações taxonômicas inicia com uma semente fornecida pelo usuário, que é uma palavra-chave do domínio para o qual a estrutura ontológica deve ser construída. Essa semente é usada para pesquisar documentos *web* relacionados ao domínio. A partir desses documentos, são extraídas relações de hiponímia as quais são usadas para construir uma taxonomia com conceitos gerais, de apenas um nível. Desses documentos também são extraídos os verbos que apareceram no contexto da semente, ou seja, na mesma sentença. Os autores usam medidas estatísticas *web* para selecionar os verbos cujas relações com a semente são mais relevantes. Os verbos escolhidos são usados em padrões de pesquisa para buscar novos textos *web*. A partir desses textos, os conceitos relacionados à semente são obtidos. Os verbos selecionados são ainda usados para nomear as relações não taxonômicas identificadas. O processo de aprendizagem se repete, usando como nova palavra-chave um dos subconceitos da semente, enriquecendo cada vez mais a estrutura conceitual com novos conceitos e relacionamentos.

Assim como Kavalec e Svátek em [104], os autores também comentam a dificuldade quanto à avaliação das relações transversais dada a ausência dessas relações em ontologias de referência. Por isso, propõem uma medida de similaridade baseada na WordNet. Eles medem, usando a medida cosseno, a relação entre dois conceitos comparando as suas glosas, como vetores de termos. Nos testes realizados, apenas 70% das relações puderam ser avaliadas, visto que seus conceitos existiam na WordNet. No entanto, a baixa cobertura da WordNet para alguns dos domínios pesquisados prejudicou fortemente os resultados e os autores optaram então por utilizar também avaliadores humanos. Estes relataram que os conceitos recuperados eram geralmente muito específicos e que da lista de verbos selecionada poucos verbos eram realmente produtivos.

Já Weichselbraum *et al.* em [216, 215] têm como objetivo, além de identificar as relações

[9] http://www.lonelyplanet.com/destinations/
[10] http://www.cs.unt.edu/~rada/downloads.html
[11] http://helmer.aksis.uib.no/icame/brown/bcm.html

não taxonômicas, atribuir às mesmas rótulos adequados. A proposta inclui o uso de uma metaontologia que descreve as relações comuns em ontologias de domínio; de recursos externos como a DBpedia[12] e a OpenCyc[13]; e de *corpora* extraídos da *web*. Esses recursos são usados para construir semiautomaticamente uma base de conhecimento com informações sobre relações conhecidas (rotuladas) entre conceitos de um determinado domínio. Para cada relação entre conceitos C_m e C_n, a base contém uma lista dos verbos que costumam aparecer em textos associando o conceito C_m a C_n; o nome (rótulo) dessa relação; uma conceituação (metaontologia) descrevendo o domínio, a imagem e as propriedades que caracterizam a relação; e ainda fragmentos de ontologias como a OpenCyc que formalizam o conhecimento sobre o domínio, descrevendo as classes dos conceitos que seguem a conceituação estabelecida para a relação. O método pode tanto estender uma ontologia incluindo novas relações quanto gerar uma estrutura conceitual a partir de palavras-chave de um determinado domínio.

Para definir o nome de uma relação desconhecida entre dois conceitos ou instâncias, pertencentes à ontologia a ser estendida ou às palavras-chave informadas, inicialmente são extraídos de textos *web*, o conjunto de verbos que aparecem nos contextos dos conceitos dessa relação. Por meio da medida cosseno, é estabelecida a similaridade entre a lista de verbos da relação desconhecida e a lista de verbos das relações presentes na base de conhecimento. A relação da base cuja similaridade for maior, terá seu rótulo atribuído à relação desconhecida. Para avaliar o método proposto, especialistas estenderam manualmente uma ontologia cujo domínio refere-se a mudanças climáticas, incluindo relações transversais. Uma parte das relações existentes nessa ontologia foi usada para construir a base de conhecimento e a parte restante, para avaliar a abordagem. Embora os resultados sejam dependentes do tipo de relação não taxonômica que está sendo avaliada, a F-*measure* ficou em torno de 84%.

Já Balakrishma *et al.* em [11] utilizam papéis semânticos para identificar e também para nomear as relações não taxonômicas. Dentre as 26 relações apresentadas pelos autores, a maioria é baseada em papéis semânticos, tais como: $Agent(X, Y)$, onde X é uma pessoa e é o agente de Y; $Instrument(X, Y)$, onde X é o instrumento de Y; $Topic(X, Y)$, onde X é o tópico de uma comunicação feita por Y; etc. O método dos autores consiste em usar classificadores conhecidos, como Árvores de Decisão, Navie Bayes e Support Vector Machine e novas propostas como Semantic Scattering [144], para reconhecer as relações nos textos. Os classificadores são treinados com informações informações morfossintáticas e semânticas extraídas de *corpora* anotados. Para avaliar a abordagem quanto à descoberta e etiquetagem correta das relações, os autores usaram, como conjunto de treino para reconhecer padrões de sintagmas nominais, os *corpora* TreeBank 2, L.A Times e XWN 2.0. Utilizaram também a FrameNet como conjunto de treino para identificar os padrões quanto aos argumentos dos verbos. Como conjunto de teste, anotaram manualmente 500 sentenças, aleatoriamente escolhidas, do TreeBank 3 com relações semânticas. Para esse conjunto de teste, os autores obtiveram 49,67 de F-*measure*.

O trabalho de Basili *et al.* em [14] também faz uso de papéis semânticos no processo de aprendizagem de ontologias a partir de texto. Os autores utilizam a WordNet para desambiguação de sentido e a FrameNet para atribuir papéis semânticos aos substantivos encontrados nos textos. Há trabalhos ainda que utilizam os papéis para caracterizar conceitos em aplicações de categorização e agrupamento de documentos [195, 194] e também para expandir consultas em aplicações de recuperação de informações [147].

4.9 Considerações sobre este capítulo

A anotação de papéis semânticos é um recurso interessante para extração e representação de estruturas ontológicas visto que permite identificar e relacionar as entidades que participam de

[12] http://dbpedia.org/
[13] http://www.opencyc.org/

um evento. No entanto, tratar com anotações semânticas desse tipo envolve pelo menos duas dificuldades: a falta de uma lista consensual de papéis semânticos e a imprecisão dos anotadores atuais para domínios que não sejam o do PropBank.

Embora o PropBank tenha, de uma certa forma ,contornado o problema da falta de consenso ao numerar os papéis, para aplicações que envolvem representação de conhecimento e venham a utilizar etiquetadores treinados com este *corpus*, o problema persiste, pois é importante determinar o significado de cada anotação. Caso a aplicação faça uso de um etiquetador cujo *corpus* de treino seja o FrameNet haverá problemas também, pois mesmo com etiquetas mais significativas, elas são específicas, ou seja, foram definidas para as situações retratadas naquele *corpus,* que talvez não sejam as mesmas de outros *corpora.*

Além disso, no caso dos etiquetadores cuja anotação segue a do PropBank, o significado das etiquetas só é padronizado para verbos que pertençam às mesmas classes semânticas. Apenas duas etiquetas $Arg0$ e $Arg1$ são mais regulares, correspondendo, respectivamente, aos papéis agente e paciente/tema na maioria das vezes.

Somado a isso ainda há o fato que a pesquisa referente ao desenvolvimento de etiquetadores automáticos de papéis semânticos é relativamente recente. Há ainda poucos *corpora* de treino mesmo para a Língua Inglesa. E os melhores F1 obtidos na tarefa de anotação para o próprio Wall Street Journal giram em torno de 80%, caindo significativamente para *corpora* de outros domínios.

Ao mesmo tempo que esses aspectos refletem os riscos de nossa pesquisa, também a tornam interessante dado que não encontramos trabalhos atuais com a nossa abordagem.

5. MATERIAIS, FERRAMENTAS E RECURSOS

Este capítulo apresenta uma descrição sucinta dos *corpora*, bases lexicais, ontologias e ferramentas utilizados em nossa pesquisa.

5.1 *Corpora*

Nesta seção são descritos brevemente os *corpora* usados nos estudos realizados durante o doutorado.

5.1.1 Penn TreeBank

O Penn TreeBank[1] é um grande *corpus* em língua inglesa, anotado com informações léxico-sintáticas que foram revisadas manualmente. É comercializado pelo Linguistic Data Consortium[2] (LDC) e possui 3 versões: TreeBank-1, TreeBank-2 e TreeBank-3.

O TreeBank-1 é a versão original deste *corpus* e possui cerca de 4,8 milhões de *tokens* provenientes de textos de diferentes fontes. Como mostra a Tabela 5.1, o TreeBank-1 contém resumos e boletins dos Departamentos de Energia e de Agricultura americanos, textos da Library of America[3], mensagens do MUC-3[4], sentenças de manuais da IBM sobre computadores, transcrições de sentenças do ATIS[5], transcrições de notícias de rádio do WBUR[6], artigos do Wall Street Journal e textos do *corpus* Brown[7] [137].

Tabela 5.1 – Descrição do TreeBank-1 [137]

Fonte	#*tokens* anotados com POS	#*tokens* anotados sintaticamente
resumos do Departamento de Energia	231.404	231.404
boletins do Departamento de Agricultura	78.555	78.555
textos da Library of America	105.652	105.652
mensagens do MUC-3	111.828	111.828
sentenças de manuais da IBM	89.121	89.121
sentenças do ATIS	19.832	19.832
artigos do Wall Street Journal	3.065.776	1.061.166
corpus Brown	1.172.041	1.172.041
transcrições de rádio WBUR	11.589	11.589
Total	4.885.798	2.881.188

Na versão TreeBank-2[8] foi introduzido um outro estilo de anotação sintática, baseado em parênteses [136] e ilustrado na Figura 5.1. Esse estilo é muito usado, atualmente. *Parsers*

[1] Mais informações pode ser encontradas no *site* do projeto Penn TreeBank: http://www.cis.upenn.edu/~treebank/

[2] http://www.ldc.upenn.edu/

[3] http://www.loa.org/

[4] 3rd Message Understanding Conferences (MUC-3), disponível em http://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html

[5] Air Travel Information Services (ATIS), mais informações em http://www.ai.sri.com/natural-language/projects/arpa-sls/atis.html.

[6] Boston University Radio Speech Corpus, disponível no catálogo da LDC

[7] http://khnt.aksis.uib.no/icame/manuals/brown/

[8] Cabe ressaltar que no pacote do TreeBank-2 distribuído pelo LDC está incluído também o TreeBank-1 com sua anotação original.

como de Stanford (seção 5.3.3) seguem essa forma de anotação. O TreeBank-2 inclui cerca de 1 milhão de *tokens*. É formado por 2.499 artigos do ano de 1989 do Wall Street Journal e uma pequena mostra de sentenças do ATIS. Inclui ainda manuais e ferramentas de anotação.

Já o TreeBank-3 contém o TreeBank-2 e informações adicionais, tais como problemas de anotação decorrentes de interrupções nas sentenças, provocadas, por exemplo, por interjeições e repetições de palavras.

Os textos são fornecidos em 4 formatos: *raw* (estado bruto), *tagged* (anotado apenas como POS), *parsed* (anotado apenas com informações sintáticas) e *combined* (marcado com etiquetas POS e sintáticas). O tipo de anotação atribuída aos textos pode ser identificado pelas extensões dos arquivos. Textos em formato bruto não possuem extensão. Os textos nos formatos *tagged*, *parsed* e *combined* possuem, respectivamente, as extensões *pos*, *prd* e *mrg*.

A Figura 5.1 ilustra o formato *combined* para a sentença em *raw*: "Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29." do texto wsj_0001.

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

Figura 5.1 – Exemplo da anotação *combined* aplicada a textos TreeBank-2

5.1.1.1 Penn TreeBank Sample

O Penn TreeBank Sample corresponde a 9% do *corpus* Treebank-2. Ele está entre os *corpora* disponibilizados livremente pela ferramenta NTLK[9]. Possui 100.673 *tokens* distribuídos em 199 textos (de wsj_0001 a wsj_0199), com uma média de 19,21 sentenças por texto.

Esta foi a versão do Penn TreeBank usada em nossos estudos iniciais. Embora seja um *corpus* relativamente pequeno, ele pode ser baixado livremente e contém um volume de relações verbo-argumento adequado para a realização de nossa pesquisa.

5.1.1.2 PropBank

Como já mencionado, o *corpus* PropBank é particionado em anotação e em léxico de verbos (*frames file*). Na versão distribuída pelo LDC, há em torno de 113.000 instâncias de verbos anotados. São instâncias de cerca de 3.300 verbos diferentes. Sendo que não há anotações para verbos de ligação como *to be* e auxilares como *to do* e *to have*.

A Figura 5.2 exibe um trecho do arquivo de anotação *prop.txt*. A primeira coluna identifica o texto e a segunda, a sentença desse texto. As sentenças são enumeradas a partir de 0 (zero). A primeira linha desse arquivo corresponde à anotação PropBank da sentença 0 do texto wsj_-0001.mrg que foi apresentada na Figura 5.1.

[9]Os corpora disponibilizados pela ferramenta NLTK podem ser encontrados em: http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml

```
wsj/00/wsj_0001.mrg 0 8 gold join.01 vf--a 0:2-ARG0 7:0-ARGM-MOD 8:0-rel 9:1-ARG1 11:1-ARGM-PRD 15:1-ARGM-TMP
wsj/00/wsj_0001.mrg 1 10 gold publish.01 p---a 10:0-rel 11:0-ARG0
wsj/00/wsj_0002.mrg 0 16 gold name.01 pp--p 16:0-rel 0:2*17:0-ARG1 18:2-ARG2
wsj/00/wsj_0003.mrg 0 5 gold use.01 p---p 4:1-ARGM-TMP 5:0-rel 0:2*6:0-ARG1 7:2-ARG2-PNC
wsj/00/wsj_0003.mrg 0 9 gold make.01 i---a 7:0-ARG0 9:0-rel 10:1-ARG1
```

Figura 5.2 – Trecho do arquivo *prop.txt* do PropBank.

Cada terminal da sentença também é enumerado a partir de zero (Figura 5.3). A terceira coluna da Figura 5.2 corresponde à posição do verbo na sentença. O verbo *to join* está na posição 8.

Pierre	Viken	,	61	years	old	,	will	join	the	board	as	a	nonexecutive	...
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	...

Figura 5.3 – Exemplo de identificação dos terminais de uma sentença PropBank.

A quarta coluna da Figura 5.2 identifica o anotador. Quando aparece a palavra *gold*, que é o caso dessa sentença, significa que mais de uma pessoa foi responsável pela anotação.

A quinta coluna identifica o *frameset* do verbo. Para esta sentença, o *frameset* é o *join.01*. Já a sexta coluna fornece 5 informações sobre a conjugação do verbo, nesta ordem: forma (*i=infinitive, g=gerund, p=participle e v=finite*), tempo (*f=future, p=past e n=present*), modo (*p=perfect, o=progressive e b=both perfect and progressive*), pessoa (*3=3rd person*) e voz (*a=active e p=passive*). A anotação *vf–a* indica que o verbo está na forma *finite*, no tempo *future* e na voz *active*. Não há informação sobre modo e pessoa.

As colunas seguintes correspondem à anotação dos papéis semânticos. A anotação *0:2-ARG0* da sétima coluna, indica que 2 níveis acima do terminal 0 (zero), além da etiqueta sintática, há também a etiqueta semântica ARG0. Na sentença 0 do texto wsj_0001.mrg, esta etiqueta corresponde ao trecho anotado como NP : Pierre Vinken (Figura 5.4).

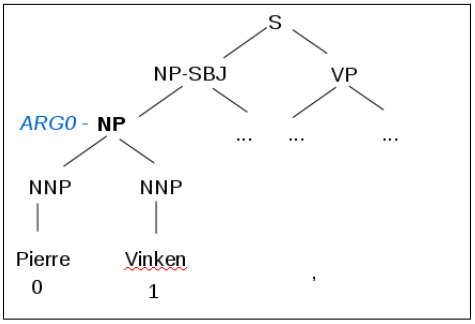


Figura 5.4 – Exemplo de anotação de papel semântico no PropBank.

Existem 4 formas de anotação para papéis semânticos. Descrevemos apenas a primeira forma. Mais informações podem ser encontradas no arquivo *readme* fornecido juntamente com o *corpus*.

Cabe ressaltar ainda que a parte referente ao léxico de verbos não foi detalhada nesta seção, pois já foi apresentada na Seção 4.3.

5.1.1.3 SemLink 1.1

O *corpus* SemLink 1.1[10] foi o que acabamos usando de fato em nossos estudos iniciais. Ele é uma extensão do *corpus* PropBank e contém informações semânticas da VerbNet. As informações

[10]Disponível em http://verbs.colorado.edu/semlink/

VerbNet adicionadas às anotações PropBank nos permitiram estudar as classes dos verbos e trabalhar com identificadores mais tradicionais para papéis semânticos. Desta forma, ao invés de apresentar em nosso estudo apenas os rótulos numéricos do PropBank, fizemos uso do mapeamento VerbNet desses rótulos a nomes de papéis semânticos mais usuais (Agent, Theme, etc).

O projeto SemLink tem como objetivo combinar diferentes recursos lexicais: PropBank, FrameNet, VerbNet e WordNet [129]. Dentro do escopo daquele projeto, nós utilizamos o mapeamento *vnpbprop.txt,* que combina o PropBank com a VerbNet. Nesse arquivo, como mostra a Figura 5.5, foram introduzidas no PropBank anotações referentes à classe do verbo e aos papéis semânticos. No caso da sentença 0 do texto wsj_0001.mrg, a classe VerbNet do verbo *to join* é 22.1-2 e os papéis semânticos PropBank ARG0 e ARG1 foram mapeados, respectivamente, em Agent e Patient1.

Como pode-se observar ainda na Figura 5.5, o mapeamento não é completo. Nem todos os papéis PropBank foram mapeados. Os papéis ARGM da sentença 0 do texto wsj_0001.mrg, por exemplo, não possuem papel semântico VerbNet correspondente. Há ainda verbos para os quais não há classe associada, como o verbo *to use* na sentença 0 do texto wsj_0003.mrg.

```
wsj/00/wsj_0001.mrg    0   8 auto     join.01;VN=22.1-2         vf--a 0:2-ARG0[Agent] 7:0-ARGM-MOD 8:0-rel 9:1-ARG1
[Patient1] 11:1-ARGM-PRD 15:1-ARGM-TMP
wsj/00/wsj_0001.mrg    1  10 vieweg   publish.01;VN=26.4-1      p---a 10:0-rel 11:0-ARG0[Agent]
wsj/00/wsj_0002.mrg    0  16 auto     name.01;VN=29.3           pp--p 16:0-rel 0:2*17:0-ARG1[Theme] 18:2-ARG2[Predicate]
wsj/00/wsj_0003.mrg    0   5 vieweg   use.01;VN=None            p---p 4:1-ARGM-TMP 5:0-rel 0:2*6:0-ARG1 7:2-ARG2-PNC
wsj/00/wsj_0003.mrg    0   9 auto     make.01;VN=26.1-1         i---a 7:0-ARG0[Agent] 9:0-rel 10:1-ARG1[Product]
```

Figura 5.5 – Trecho do arquivo *vnpbprop.txt* do SemLink 1.1.

Como usamos apenas os 199 textos (de wsj_0001 a wsj_0199) do SemLink que correspondem ao Penn TreeBank 2, chamaremos esse subconjunto de SemLink 1.1 Sample.

5.1.2 Wikicorpus 1.0

O Wikicorpus[11] 1.0 é um *corpus* trilíngue, com textos Wikipédia em Catalão, Espanhol e Inglês. Esses textos foram selecionados e filtrados por Reese *et al.*, a partir de um *dump* de 2006 da Wikipédia. Possui cerca de 750 milhões de palavras, sendo que 600 milhões correspondem à Lingua Inglesa. As palavras foram anotadas automaticamente com informações linguísticas [178].

Este *corpus* pode ser baixado livremente e fornece os artigos em 2 formatos: *raw* (estado bruto) e *tagged* (com anotações linguísticas). A Figura 5.6 ilustra o formato *tagged*. No cabeçalho dos textos, há informações sobre o artigo (identificadores e título). As linhas posteriores estão organizadas em 4 colunas: a primeira é o *token*; a segunda, seu lema; a terceira, seu POS e a quarta, seu sentido WordNet. Quando o sentido WordNet não pode ser identificado, este código aparece como zero.

Em nossos estudos, usamos um subconjunto de textos desse *corpus*. Como precisavámos de anotações referentes a papéis semânticos, as quais não estavam presentes na versão *tagged*, utilizamos a versão em formato *raw*. Os textos, em formato *raw*, presentes nesse subconjunto, foram anotados semanticamente pela ferramenta F-EXT-WS descrita na Seção 5.3.4.

Para realizarmos nossa pesquisa, precisávamos de *corpora* de domínios específicos. Julgamos que o domínio Finanças fosse o mais adequado, visto que poderia maximizar o desempenho dos etiquetadores de papéis semânticos. A maioria dessas ferramentas, como a F-EXT-WS, têm sido treinadas usualmente com o PropBank, cujos textos são de Finanças. Para complementar nossa investigação, precisávamos ainda de um segundo domínio. Selecionamos, então, o domínio Turismo, por ser uma escolha frequente em trabalhos relacionados ao nosso [13, 41].

[11] Disponível para download em http://nlp.lsi.upc.edu/wikicorpus/

```
<doc id="1881893" title="Monster-in-Law"
nonfiltered="39" processed="39" dbindex="1480038">
...
It it PRP 0
marks mark VBZ 0
a 1 Z 0
return return NN 03235693
to to TO 0
cinema cinema NN 04735661
for for IN 0
Jane_Fonda jane_fonda NNP 0
; ; Fx 0
her her PRP$ 0
first 1 JJ 02098880
film film NN 04960631
in in IN 0
15 15 Z 0
years year NNS 0
. . Fp 0
...
</doc>
```

Figura 5.6 – Exemplo de anotação do WikiCorpus (Inglês)

Como no cabeçalho dos documentos do Wikicorpus 1.0 não há informação de categoria, para que pudéssemos selecionar artigos dos domínios mencionados, foi necessário buscar essa informação na própria Wikipédia. Para isso, processamos um *dump* de 20/01/2011 e construímos uma lista com todos os títulos de artigos das categorias *Finance* e *Financial* para identificar textos do domínio Finanças. Para o domínio Turismo fizemos o mesmo, geramos uma outra lista contendo todos os títulos de artigos das categorias *Tourism* e *Travel.* Os *corpora* formados a partir dessas listas foram chamados de WikiFinance e WikiTourism e são comentados, respectivamente, nas Seções 5.1.2.1 e 5.1.2.2.

É importante ressaltar que dos 8.958 artigos extraídos do Wikicorpus 1.0, utilizamos em nossa pesquisa apenas 934. Esses 8.958 textos, no entanto, não correspondem ao total de artigos do Wikicorpus 1.0, mas sim à quantidade de textos que conseguimos processar. O arquivo compactado do Wikicorpus 1.0 que está disponível na *web* é formado por 164 arquivos-texto. Cada um desses arquivos-texto contém vários artigos Wikipédia. Desses 164, 9 foram baixados vazios (com 0 byte) e 1, processado parcialmente. Isso corresponde a uma perda de aproxidamente 6%. Investimos algum tempo na resolução desses problemas, no entanto acabamos decidindo por aceitar a perda, e trabalhando apenas com um subconjunto dado que já era suficiente para os propósitos de nossa pesquisa.

Para viabilizar o uso de etiquetadores sintáticos e semânticos nos textos selecionados, foi necessário ainda eliminar *tags* e trechos que poderiam prejudicar o desempenho desses anotadores. Excluímos, por exemplo, *tags* de identificação (*id, title*, ...) e trechos relativos a *links* para outras páginas *web*, os quais são muito comuns em textos Wikipédia. O processo de identificação de categorias, limpeza e formação desses *corpora* foi realizado a partir de programas em linguagem C que implementamos para esse fim.

Cabe mencionar, por fim, que escolhemos a Wikipédia por conter artigos mais conceituais, o que julgamos mais adequado quando a finalidade é a extração de informações para construção de estruturas ontológicas.

5.1.2.1 WikiFinance

O WikiFinance é um subconjunto do *corpus* Wikicorpus 1.0 para o Inglês e é constituído por 482 textos do domínio Finanças. Ele possui 193.836 *tokens* distribuídos em 945 sentenças. A quantidade de sentenças presentes nos textos não é uniforme. Na média, há ao menos 2 sentenças por texto.

Inicialmente esses textos foram anotados com o Stanford *parser* 1.6.5 (Seção 5.3.3). Este *parser* nos proveu anotações lexicossintáticas e de dependências. No entanto, por questões de simplicidade acabamos não usando tais anotações. Evitamos desta forma a necessidade de alinhamento das anotações realizadas pelo Stanford *parser* com as da ferramenta F-EXT-WS (Seção 5.3.4), usada para marcar os papéis semânticos. Acabamos utilizando as anotações lexicossintáticas providas pela própria F-EXT-WS.

A partir dessas últimas anotações analisamos a distribuição dos tokens do *corpus* WikiFinance em categorias lexicais, contabilizando suas etiquetas POS. A Tabela 5.2 apresenta a frequência das categorias relativas a substantivos, adjetivos e verbos, que são mais relevantes para o nosso estudo.

Tabela 5.2 – Quantidade de substantivos, adjetivos e verbos existentes no WikiFinance

Classe Gramatical (etiqueta POS)	#
Substantivos Comuns (NN e NNS)	47.276
Verbos (VB, VBD, VBZ, VBG, VBN, VBP)	26.844
Substantivos Próprios (NNP e NNPS)	17.939
Adjetivos (JJ)	14.748

5.1.2.2 WikiTourism

O WikiTourism também é um subconjunto do *corpus* Wikicorpus 1.0 para o Inglês e contém 442 textos do domínio Turismo. Ele possui 179.399 *tokens* distribuídos em 824 sentenças. Da mesma forma que o WikiFinance, a quantidade de sentenças presentes nos textos não é uniforme. Contém, em média, cerca de 2 sentenças por texto.

A partir das anotações providas pela ferramenta F-EXT-WS (Seção 5.3.4), também analisamos a distribuição dos *tokens* do *corpus* WikiTourism nas categorias lexicais referentes a substantivos, adjetivos e verbos. A Tabela 5.3 apresenta tal distribuição.

Tabela 5.3 – Quantidade de substantivos, adjetivos e verbos existentes no WikiTourism

Classe Gramatical (etiqueta POS)	#
Substantivos Comuns (NN e NNS)	36.098
Verbos (VB, VBD, VBZ, VBG, VBN, VBP)	21.772
Substantivos Próprios (NNP e NNPS)	27.174
Adjetivos (JJ)	12.416

5.1.3 PLN-BR CATEG

O corpus PLN-BR CATEG[12] foi utilizado em estudos para Língua Portuguesa. Ele contém cerca de 30 mil textos (9.780.220 tokens) do jornal Folha de São Paulo entre os anos de 1994 e

[12] O *corpus* PLN-BR CATEG foi constituído a partir do projeto "Recursos e Ferramentas para Recuperação de Informações em Bases Textuais em Português do Brasil" (PLN-BR) financiamento CNPq - CT INFO #550388/2005-2. Mais informações sobre o projeto podem ser encontradas em http://www.nilc.icmc.usp.br/plnbr/

2005.

Os textos desse *corpus* estão organizados em 29 seções: Agrofolha (193 textos), Brasil (5.606 textos), Caderno Especial (509 textos), Caderno Especial 2 (50 textos), Ciência (182 textos), Construção (7 textos), Cotidiano (6.458 textos), Dinheiro (4.153 textos), Empregos (238 textos), Entrevistada 2 (4 textos), Equilíbrio (28 textos), Esporte (4.632 texts), Folha Invest (165 textos), Folha Negócios (36 textos), Folha Sinapse (11 textos), Folha Teen (260 textos), Folhinha (78 textos), Folha Vest (82 textos), Ilustrada (2.935 textos), Imóveis (120 textos), Informática (408 textos), Mais! (252 textos), Mundo (2.410 texts), Primeira Página (170 textos), Revista da Folha (3 textos), Tudo (95 textos), Turismo (464 textos), TVFolha (236 textos) e Veículos (215 textos).

Para que pudessemos utilizá-lo, submetemos esse *corpus* ao lematizador FORMA, mencionado na Seção 5.3.6.

5.2 Bases lexicais e ontologias

Usamos em nosso estudo as bases lexicais VerbNet e WordNet, já comentadas, respectivamente, nas Seções 4.5 e 4.6 do capítulo anterior. Ambas foram acessadas a partir do pacote NLTK (Seção 5.3.2). Por meio desse pacote, conseguimos buscar classes VerbNet e também utilizar as medidas de similaridade semântica baseadas na WordNet. Mais especificamente, usamos a implementação, disponível no pacote, da medida de Wu e Palmer [224] no cálculo da coesão lexical dos conceitos formais. Para esse fim foi usada a versão WordNet 3.0.

Também com o propósito de calcular a coesão lexical dos conceitos e ainda para validar nossos estudos em categorização de textos, usamos ontologias do domínio de Finanças e Turismo. Essas ontologias são para Língua Inglesa e estão representadas em OWL.

Para o domínio de Finanças, encontramos 3 ontologias: SUMO_Finance[13], LSDIS_Finance[14] e Finance[15]. Tais ontologias possuem conceitos sobre instrumentos financeiros, partes envolvidas, processos e procedimentos relacionados a títulos. Tanto a ontologia SUMO_Finance quanto LSDIS_Finance foram criadas pelo grupo de pesquisa Large Scale Distributed Information System (LSDIS) do departamento de Ciência da Computação da Universidade da Georgia. Ambas possuem relações transversais, sendo que a LSDIS_Finance é uma extensão da SUMO_Finance. E foi por esta razão que, dessas duas, apenas a LSDIS_Finance foi escolhida para o uso em nossos estudos iniciais. Já a ontologia Finance, desenvolvida por Eddy Vanderlinden, foi encontrada no repositório de ontologias do Protégé[16].

Ainda no repositório de ontologias do Protégé, encontramos a ontologia Travel[17] do domínio de Turismo, criada por Holger Knublauch. Também com o nome de Travel[18], encontramos outra ontologia desse domínio. Esta última foi criada por Danica Damljanovic e faz parte de um sistema *web* na área de Turismo que este autor desenvolveu. A ontologia de Damljanovic é uma extensão da ontologia PROTON Upper Module, também chamada de WORLD, a qual define conceitos relativos a locais, datas, línguas, etc. Ambas as ontologias Travel possuem conceitos relacionados a pacotes de viagens, tipos de viajantes, destinos turísticos, etc. Para evitar confusões entre as ontologias homônimas, chamaremos a última de TGPROTON, que é o nome do seu arquivo *owl*.

Na Tabela 5.4 apresentamos dados gerais das ontologias utilizadas em nossa investigação. Esses dados são provenientes do processamento que realizamos em tais estruturas.

[13] http://lsdis.cs.uga.edu/projects/meteor-s/wsdl-s/ontologies/SUMO_Finance.owl
[14] http://lsdis.cs.uga.edu/projects/meteor-s/wsdl-s/ontologies/LSDIS_Finance.owl
[15] http://www.fadyart.com/ontologies/data/Finance.owl
[16] http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library
[17] http://protege.cim3.net/file/pub/ontologies/travel/travel.owl
[18] http://goodoldai.org/ns/tgproton.owl

Tabela 5.4 – Dados gerais das ontologias utilizadas em nosso estudo

Nome	**Domínio**	**#Classes**	**Altura**
LSDIS_Finance	Finanças	270	7
Finance	Finanças	323	7
Travel	Turismo	34	5
TGPROTON	Turismo	58	4

Cabe mencionar que apesar de existirem muitas ontologias na *web*, nem sempre há informações suficientes que nos permitam avaliar concretamente a relevância de tais ontologias para os domínios aos quais elas se propõem. Obviamente, as ontologias com mais classes, como as de Finanças, aparentam maior relevância dado o maior detalhamento do domínio. No entanto, quantidade em classes não corresponde necessariamente à qualidade em conceitos.

Dada a falta de informação de ordem semântica dessas ontologias, baseamos nossas escolhas na procedência dessas estruturas. Logo, consideramos relevantes as ontologias existentes em repositórios conhecidos, como o do Protégé, e aquelas desenvolvidas e disponibilizadas por pesquisadores.

5.3 Ferramentas

Nesta seção descrevemos brevemente as ferramentas utilizadas no pré-processamento dos textos e em tarefas de agrupamento.

5.3.1 TreeTagger

O TreeTagger[19] é um lematizador e etiquetador de POS. Ele tem sido usado em várias pesquisas, principalmente porque atende a várias línguas, tais como inglês, alemão e português.

Decidimos usá-lo, pois em alguns testes empíricos de lematização dos *corpora* Penn TreeBank Sample e WikiFinance, observamos que o TreeTagger teve mais sucesso que o lematizador do pacote NLTK (Seção 5.3.2). Os erros mais significativos ocorreram com verbos, como mostra a Tabela 5.5.

Tabela 5.5 – Erros de lematização

Termo original	lematizador NLTK	lematizador Treetagger
'm	'm	be
're	're	be
's	's	be
've	've	have
annualized	annualized	annualize

5.3.2 Pacote NLTK

O Natural Language Toolkit[20] (NLTK) é um pacote de ferramentas de código aberto, escrito em Python, usado para o processamento da linguagem natural. O pacote inclui tokenizadores, *stemmers*, lematizadores, *chunkers*, *parsers*, clusterizadores e classificadores [20]. Junto com as ferramentas são disponibilizadas também amostras de *corpora*, tais como Brown, Reuters, e

[19] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[20] http://www.nltk.org/

Penn TreeBank. O pacote ainda contém ferramentas que permitem o acesso às bases lexicais como a WordNet.

Escolhemos este pacote (versão nltk-2.0b9) justamente por reunir diferentes ferramentas essenciais para processamento de textos, das quais usamos efetivamente o *stemmer* e as interfaces para WordNet e VerbNet.

Inicialmente, estavamos usando também o lematizador disponível no pacote, mas em razão de erros na lematização de verbos, optamos pelo uso da ferramenta Tree Tagger (Seção 5.3.1) para esse fim, como já exposto.

Cabe mencionar que para usar cada uma das ferramentas citadas, tivemos que implementar pequenos programas em Python.

5.3.3 Stanford Parser

O *parser* estatístico da universidade de Stanford anota informações lexicossintáticas ao estilo TreeBank-2 e também relações de dependência lexical. O *parser* foi criado para marcar textos em Língua Inglesa, mas pode ser adaptado para realizar anotações em outras línguas [138].

Escolhemos este *parser* por poder ser obtido e usado livremente e, principalmente, por anotar relações de dependência, que facilitam a identificação dos elementos essenciais à nossa pesquisa, que são os verbos e seus argumentos. Além disso, provê anotação ao estilo TreeBank-2 o que também nos favorece, pois evita que tenhamos que trabalhar com anotações em formatos diversos. Sem contar ainda que é um *parser* bem conceituado cuja precisão é em torno de 86% para anotação sintática e de 91% para anotação de dependências [111].

Usamos a versão mais recente do *parser*[21] para Língua Inglesa, 1.6.5, escrita em Java, para anotar os *corpora* Penn TreeBank Sample e WikiFinance.

A anotação sintática e a anotação de dependência são geradas separadamente, ou seja, o *parser* devolve duas saídas para o mesmo texto. Na Figura 5.7, é apresentado um exemplo de anotação para a sentença "Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29." do texto wsj_0001 do *corpus* Penn TreeBank.

```
(ROOT                                                                  
  (S                                                                   
    (NP                                                                
      (NP (NNP Pierre) (NNP Vinken))                    nn(Vinken-2, Pierre-1)
      (, ,)                                             nsubj(join-9, Vinken-2)
      (ADJP                                             num(years-5, 61-4)
        (NP (CD 61) (NNS years))                        npadvmod(old-6, years-5)
        (JJ old))                                       amod(Vinken-2, old-6)
      (, ,))                                            aux(join-9, will-8)
    (VP (MD will)                                       det(board-11, the-10)
      (VP (VB join)                                     dobj(join-9, board-11)
        (NP (DT the) (NN board))                        det(director-15, a-13)
        (PP (IN as)                                     amod(director-15, nonexecutive-14)
          (NP (DT a) (JJ nonexecutive) (NN director)))  prep_as(join-9, director-15)
        (NP (NNP Nov.) (CD 29))))                       tmod(join-9, Nov.-16)
    (. .)))                                             num(Nov.-16, 29-17)
```

Figura 5.7 – Exemplo de anotação sintática e de dependência feita pelo Stanford *parser*

Como já mencionado, por razões de simplicidade, acabamos não usamos os textos anotados por este *parser*. Utilizamos as anotações linguísticas providas pelo processador F-EXT-WS, que é apresentado na próxima seção.

[21] http://nlp.stanford.edu/software/lex-parser.shtml

5.3.4 Etiquetadores de Papéis Semânticos

Para anotar os textos com papéis semânticos, utilizamos duas ferramentas: Illinois Semantic Role Labeler[22] (Illinois SRL) e F-EXT-WS[23]. A ferramenta Illinois SRL foi utilizada de forma mais superficial, em estudos preliminares para Língua Portuguesa. Já a ferramenta F-EXT-WS foi responsável pela anotação semântica dos *corpora* WikiFinance e WikiTourism, utilizados em nossos estudos para a tarefa categorização de textos. Cabe ressaltar que ambas as ferramentas utilizam rótulos PropBank para etiquetar os papéis semânticos e disponibilizam esse tipo de anotação apenas para textos em Língua Inglesa.

A ferramenta Illinois SRL foi desenvolvida pelo Cognitive Computation Group (CGC) da Universidade de Illinois [174]. Possui uma versão demo[24] na *web* e uma versão que pode ser usada remotamente a partir da instalação de um programa cliente. Esse programa, no entanto, depende de outros recursos que incluem bibliotecas e *parsers*. Encontramos dificuldades quanto à configuração correta dos recursos necessários ao seu funcionamento. Como não conseguimos resolver tais problemas no tempo destinado para a realização dessa tarefa, dentro do cronograma de nossa pesquisa, desistimos de utilizá-lo. Para os estudos preliminares em Língua Portuguesa, usamos a versão demo.

Nos estudos com *corpora* para Língua Inglesa, utilizamos a ferramenta F-EXT-WS. Ela é um processador *web* para linguagem natural [60]. Foi desenvolvida pelo grupo de pesquisa LEARN do departamento de Ciência da Computação da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). O processador F-EXT-WS oferece serviços de anotação tanto para textos em Língua Portuguesa quanto Inglesa, sendo que a anotação semântica é disponibilizada apenas para esta última. Para Língua Inglesa, ele provê anotações para POS, sintagmas, cláusulas e papéis semânticos. Para utilizá-lo, basta submeter o texto em formato *raw,* escolher o tipo de anotação desejada, que a ferramenta gera um outro arquivo com as anotações solicitadas.

Em nossos estudos usamos todas as anotações providas pela ferramenta. A Figura 5.8 mostra um exemplo da anotação realizada pelo F-EXT-WS. No exemplo, o anotador processou a sentença: "Lockbox is generally divided into Wholesale and Retail.", do *corpus* WikiFinance. Esta sentença pertence ao texto wiki_9671446.txt cujo título é "Lock box".

```
[features = word, pos, ck, clause, verb, srl0, srl1, srl2, srl3] [taggingTime=00:00:00]

Lockbox          NNP  B-NP   (S*    -        (A1*)       *     *     *
is               VBZ  B-VP   *      -        *           *     *     *
generally        RB   I-VP   *      -        (AM-ADV*)   *     *     *
divided          VBN  I-VP   *      divide   (V*)        *     *     *
into             IN   B-PP   *      -        (A2*        *     *     *
Wholesale        JJ   B-NP   *      -        *           *     *     *
and              DT   I-NP   *      -        *           *     *     *
Retail           JJ   I-NP   *      -        *)          *     *     *
.                .    O      *S)    -        *           *     *     *
```

Figura 5.8 – Exemplo de anotação do F-EXT-WS

Observando-se a figura, pode-se perceber que a primeira coluna refere-se ao termo (*word*); a segunda, a anotação de POS; a terceira, à anotação de sintagma (*ck*); a quarta, à identificação de sentenças (*clause*); a quinta, ao verbo e as demais, aos papéis semânticos (*srl0*, *srl1*, *srl2* e *srl3*). Para a sentença do exemplo, o F-EXT-WS associou aos argumentos "Lockbox", "generally" e "into Whosale and Retail" do verbo *to divide*, respectivamente, os papéis semânticos A1, AM-ADV e A2.

[22] http://cogcomp.cs.illinois.edu/page/software_view/SRL
[23] http://www.learn.inf.puc-rio.br/fextws/index.jsf
[24] http://l2r.cs.uiuc.edu/~cogcomp/srl-demo.php

Com o objetivo de termos uma referência quanto ao desempenho de ambos os etiquetadores de papéis semânticos, pesquisamos suas medidas de precisão, *recall* e F1 quanto à tarefa de anotação semântica solicitada na CoNLL-2005 (Conference on Computational Natural Language Learning do ano de 2005). A tarefa, nesse ano da conferência, consistia em marcar com papéis semânticos textos do Wall Street Journal. Para o referido *corpus*, o anotador Illinois SRL obteve 82,28% na medida precisão; 76,78% na *recall* e 79,44% na F1 [174]. Já o processador F-EXT-WS atingiu o índice de 80,54% na precisão; 65,47% na *recall* e 72,23% na F1 [188]. É importante ressaltar que as fontes das quais extraímos esses dados não são recentes, e, por isso, é bem possível que ambos os anotadores já tenham melhorado os seus desempenhos.

Por fim, cabe mencionar que, para extrair as anotações providas pelo F-EXT-WS, desenvolvemos um *parser* em linguagem Java.

5.3.5 Ferramentas para gerar FCA e extensões

Usamos duas ferramentas para gerar as representações gráficas das estruturas conceituais do tipo FCA apresentadas neste documento. São elas: Concept Expert 1.3 e Eclipse's Relational Concept Analysis (ERCA).

A ferramenta Concept Expert 1.3[25], desenvolvida pela equipe de Serhiy Yevtushenko, foi utilizada, ao longo do texto, na maioria das figuras que ilustram estruturas FCA. Optamos por utilizá-la por ser uma ferramenta de fácil instalação e uso, além de ser frequentemente citada em trabalhos científicos que fazem uso do método FCA.

Já a ferramenta ERCA[26] foi usada para gerar as estruturas do tipo RCA. Essa ferramenta foi desenvolvida pelo Laboratório de Informática, Robótica e Microeletrônica de Montpellier (LIRMM). Esse laboratório reúne pesquisadores de duas instituições: Universidade de Montpellier 2 e Centro Nacional de Pesquisas Científicas. Ela foi a única ferramenta que encontramos capaz de gerar estruturas RCA.

Cabe mencionar que os grafos correspondentes às estruturas FCA, analisadas em nossos estudos, foram gerados a partir do algoritmo Bordat (Seção 3.5.2), o qual implementamos em linguagem Java.

5.3.6 Ferramentas usadas em estudos para a Língua Portuguesa

Utilizamos no pré-processamento do *corpus* PLN-BR CATEG (Seção 5.1.3), as ferramentas FORMA e WordSmith Tools. Usamos, ainda, em estudos com este mesmo *corpus* a ferramenta Clustering ToolKit (CLUTO), no entanto, para construir conceitos a partir de termos extraídos dos textos.

A ferramenta FORMA foi desenvolvida por Marco Gonzalez durante seu doutoramento. Esta ferramenta segmenta o texto, lematiza e atribui etiquetas morfológicas para palavras e sinais de pontuação, com precisão em torno de 95% [81]. A precisão inclusive foi uma das razões para a escolha dessa ferramenta. A outra razão foi o fato de a ferramenta FORMA ter sido desenvolvida por um integrante do grupo de pesquisa do qual fazemos parte, o que facilitava o suporte quanto ao uso do lematizador.

Já as outras duas ferramentas, escolhemos por serem amplamente conhecidas. O WordSmith[27], desenvolvido por Mike Scott, é uma ferramenta que auxilia na identificação de padrões textuais. Usamos uma versão demo dessa ferramenta para encontrar palavras-chave para as categorias (seções) do *corpus* PLN-BR CATEG.

[25] http://sourceforge.net/projects/conexp/
[26] http://code.google.com/p/erca/
[27] http://www.lexically.net/wordsmith/index.html

Para formar conceitos a partir de termos extraídos da seção Esportes do *corpus* PLN-BR CATEG, utilizamos alguns dos algoritmos de agrupamento disponíveis na ferramenta CLUTO[28]. Esta ferramenta foi desenvolvida pelo professor George Karypis do departamento de Ciência da Computação e Engenharia da Universidade de Minnesota. O CLUTO é uma ferramenta de livre utilização, com muitos recursos e ampla documentação.

[28] http://glaros.dtc.umn.edu/gkhome/views/cluto

6. ESTUDOS REALIZADOS - PREÂMBULO E EXTRAÇÃO DE INFORMAÇÕES DE *CORPORA* EM LÍNGUA INGLESA

Este capítulo faz uma introdução à investigação realizada neste trabalho, destacando os objetivos traçados e os métodos utilizados para alcançá-los. São apresentados trabalhos que serviram de inspiração e de base para proposta, a questão de pesquisa e seus desdobramentos, os métodos de pesquisa e avaliação empregados, bem como os estudos realizados para extração de informações de *corpora* em Língua Inglesa.

6.1 Trabalhos relacionados: diferenças e semelhanças

Combinar estruturas FCA com papéis semânticos não é uma ideia nova. Kamphuis e Sarbo em [101], na década de 90, propõem a representação de uma frase em linguagem natural, associando esses elementos. A ideia dos autores era viabilizar a interpretação de textos organizando as frases em estruturas conceituais. O método FCA foi usado para representar hierarquicamente as relações linguísticas presentes nas frases. Kamphuis e Sarbo, neste artigo, trabalharam com dois tipos de relações linguísticas: *minor* e *major*. *Minor*, tipicamente, relacionava substantivos a adjetivos e advérbios; e *major*, verbos a substantivos. Nesse estudo foram associados manualmente papéis semânticos aos substantivos contidos nas relações *minor* e *major*.

Apesar de a abordagem parecer promissora na época em que foi proposta, até agora havia sido pouco explorada. Provavelmente devido à dificuldade de anotação dos textos, visto que o surgimento de etiquetadores automáticos de papéis semânticos é mais recente. Nossa abordagem não se assemelha à de Kamphuis e Sarbo nem no método e nem no propósito. Nosso estudo, diferente do trabalho dos autores, extrai automaticamente as relações a partir de textos com anotações linguísticas. Além disso, restringe-se às relações que os autores chamam de *major*. Igualmente não visa a interpretação de textos, apenas a tangencia, na medida em que objetiva a construção automática de estruturas conceituais.

Na década de 90, Rudolf Wille em [219] apresenta exemplos de estruturas FCA combinadas com papéis semânticos. O objetivo do autor, no entanto, é diferente do nosso. Ele combinou grafos conceituais com estruturas FCA visando a formalização de uma lógica útil à representação e ao processamento de conhecimento. Nos exemplos apresentados pelo autor, os grafos conceituais, que são mapeados em estruturas FCA, contêm papéis semânticos associados. Esses papéis aparecem como atributos nos contextos formais dessas estruturas. Os contextos são formados basicamente por relações *objeto-atributo* do tipo *instância-papel_semântico* e *instância-classe*. Algumas das relações usadas em nosso estudo foram inspiradas nas definidas por Wille. Como não há comentários, naquele trabalho, quanto ao processamento das informações presentes nos grafos conceituais, imaginamos que nem a construção desses grafos e nem o seu mapeamento em estruturas FCA tenham sido realizados de forma automática. Esta, portanto, é mais uma diferença em relação ao nosso estudo. O trabalho Rudolf Wille [219] não convive com as dificuldades de automaticamente extrair informações de texto para gerar estruturas de representação de conhecimento e nem tampouco analisa, nesse sentido, os limites de sua abordagem.

Em trabalhos mais recentes também encontramos o método FCA combinado com papéis semânticos. Um exemplo é o trabalho de Valverde-Albacete [210], publicado em 2008. De forma distinta de nosso trabalho, o autor não usa o FCA como método de apoio à construção de estruturas ontológicas a partir de textos. Seu esforço é voltado à análise linguística tendo como propósito representar a FrameNet através de reticulados conceituais. Sendo assim, não faz uso, como nós, de informações textuais e nem faz uso de anotações PropBank para identificar

os papéis.

Há ainda as aplicações descritas na Seção 3.9 que, embora não utilizem papéis semânticos, estão relacionadas ao nosso estudo. Elas utilizam abordagens centradas em verbos para extrair relações não taxonômicas e pesquisam o uso do FCA ou de suas extensões como métodos de agrupamento para construção de estruturas conceituais a partir de textos. Dentre esses trabalhos, o de Cimiano *et al.* [41] é o que mais se aproxima de nossa proposta. No entanto, aqueles autores não consideram, em sua abordagem, a ambiguidade existente nas relações entre verbos e seus argumentos, nem tampouco consideram os diferentes papéis semânticos que cada argumento pode assumir ao longo de um texto.

As aplicações descritas na Seção 4.8, ainda que não façam uso do FCA como método de agrupamento, têm igualmente relação com a nossa proposta. O aspecto comum refere-se ao fato de usarem verbos e papéis semânticos na definição de relações transversais entre conceitos.

Mesmo com a profunda revisão bibliográfica realizada, não encontramos, até o momento, trabalhos que explorem os papéis semânticos em conjunto com o método FCA para apoiar a construção de estruturas ontológicas a partir de textos. Dado que a proposta de combiná-los é pouco pesquisada, consideramos ser de interesse todo esforço gerado ao explorá-la. No entanto, acreditamos que uma de nossas principais contribuições esteja no estudo de como usar a informação semântica provida pelos papéis semânticos em estruturas conceituais geradas a partir do método FCA.

A seguir, apresentamos nossa questão de pesquisa e seus desdobramentos.

6.2 Questão de Pesquisa

Como já mencionado na Seção 2.5.2, a aprendizagem de relações não taxonômicas a partir de textos é ainda um desafio [104, 134, 187]. A captura de relações dessa natureza é difícil até mesmo para os ontologistas, dada a quantidade e variedade de relacionamentos possíveis entre os conceitos [104]. A falta de "padronização" de rótulos para tais relações é um dos fatores que tem restringindo, inclusive, a aplicação das ontologias em ambientes mais dinâmicos [216].

Com o objetivo de auxiliar os ontologistas, diferentes abordagens computacionais têm sido propostas, tanto para identificar as relações quanto para nomeá-las. A maioria das abordagens pesquisadas baseia-se fundamentalmente em verbos para realizar tais tarefas.

Nossa abordagem também segue nessa linha pois, assim como Navok e Hearst em [150], acreditamos que os verbos consigam capturar aspectos sutis de significado sendo, portanto, importantes fontes de expressividade em tarefas de representação semântica. Entendemos que essa expressividade se estenda também às informações semânticas providas por verbos, como classe de verbos e papéis semânticos. Acreditamos que tais informações possam aumentar a relevância semântica de estruturas conceituais geradas a partir de textos, principalmente no que se refere à extração e à atribuição de nomes a relações não taxonômicas.

Desta forma, nosso objetivo primário é combinar o método FCA com papéis semânticos para construir, de forma automática e a partir de informações textuais, estruturas ontológicas fortemente baseadas em relações não taxonômicas. A idéia é explorar as vantagens do FCA como um método de agrupamento conceitual [171]. O método FCA, quando comparado a outros métodos de agrupamento, permite delinear mais facilmente, do ponto de vista semântico, os grupos e subgrupos de uma hierarquia [98]. Já no que tange aos papéis semânticos, investigamos o uso dessa informação semântica no processo de extração de informação. Sendo que nosso objetivo principal é utilizar tal informação na geração de conceitos (agrupamento), bem como na definição de relacionamentos entre esses conceitos.

Considerando todos esses aspectos, direcionamos nossa pesquisa por um caminho que nos permitisse analisar e entender a real contribuição dos papéis semânticos na construção de estruturas conceituais FCA geradas a partir de texto. Esse caminho de pesquisa se desdobrou

em muitas indagações as quais envolvem aspectos mais específicos referentes a:

- identificação das relações não taxonômicas: Quais os papéis semânticos mais interessantes para a identificação de relações não taxonômicas? Para que tipo de relações tais papéis são mais adequados? Apenas para as dependentes de domínio? Os papéis semânticos podem ser usados para nomear as relações não taxonômicas? Visto que alguns autores usam verbos para rotular essas relações, as classes de verbos não poderiam ser usadas para classificar tais grupos de relações ?

- organização dessas relações em uma estrutura conceitual: Considerando que o FCA é um método de agrupamento e é capaz de organizar os conceitos em uma hierarquia, como incluir a informação acerca dos papéis semânticos visto que eles provêem relacionamentos transversais entre os conceitos? RCA, que é uma extensão de FCA, é de fato mais adequado para representar a informação provida pelos papéis semânticos?

- avaliação da abordagem proposta: Dados os problemas inerentes à avaliação de estruturas conceituais geradas a partir de textos, principalmente no que tange a relações transversais, como avaliar de forma automática nossa proposta e os resultados obtidos ? O uso de ontologias de referência é viável ? Outras métricas baseadas na WordNet, como a definida por Sánchez e Moreno em [187], podem ser usadas?

- aplicabilidade da abordagem em outros *corpora* e domínios: Visto que, em sua maioria, os etiquetadores de papéis semânticos atuais seguem a anotação PropBank e que os rótulos númericos providos por esta anotação, em geral, só podem ser alinhados quando os verbos que estabelecem os papéis pertencem à mesma classe de verbos, como aplicar nosso estudo à construção de estruturas conceituais geradas a partir de outros *corpora* de diferentes domínios? Quais as limitações de nossa abordagem ?

Essas indagações são retomadas na Seção 6.3, que descreve os métodos de pesquisa usados neste trabalho e apresenta os estudos realizados durante nossa investigação.

É importante destacar que nosso estudo não gera, como resultado, ontologias tais como descritas na visão de Guarino [86], mas sim estruturas conceituais que podem ser usadas para construir ontologias. Por outro lado, nossas estruturas são ontológicas e de domínio, e suas características cabem perfeitamente na definição de Gruber [85] para ontologias, que é aceita e utilizada no âmbito dessa pesquisa.

A seguir, descrevemos os métodos de pesquisa usados para responder às questões colocadas.

6.3 Métodos de pesquisa

Com o objetivo de responder às questões colocadas na Seção 6.2, usamos os quatro aspectos apresentados naquela seção para orientar nosso estudo. Desta forma, para analisar o aspecto referente à "identificação das relações não taxonômicas", realizamos uma pesquisa exploratória com um caráter mais quantitativo. Nosso objetivo, nessa fase, era identificar, nos *corpora* usados para estudo inicial, os sintagmas nominais, os papéis semânticos, as relações e as classes de verbos que nos poderiam prover uma pesquisa mais expressiva do ponto de vista quantitativo. Chamamos a esta fase de "Extração e análise quantitativa preliminar das informações existentes nos *corpora* Penn TreeBank Sample e SemLink 1.1" e a descrevemos na Seção 6.4.

Para o aspecto referente à "organização dessas relações em uma estrutura conceitual" realizamos primeiramente uma pesquisa de natureza qualitativa. Estudamos a adequação de estruturas RCA no que se refere à representação dos papéis semânticos em conceitos formais. Analisamos, também de forma qualitativa, duas classes VerbNet. Essas classes foram apontadas pela pesquisa quantitativa, descrita na Seção 6.4, como as mais significativas para os *corpora*

analisados. Esse estudo é o assunto do Capítulo 7, ao qual chamamos de "Estudo I - Análise de estruturas conceituais RCA e de classes de verbos".

Ainda para o aspecto relativo à organização das relações, realizamos uma pesquisa exploratória. Nosso objetivo foi propor formas de incluir os papéis semânticos em reticulados conceituais, baseados em FCA, e analisar a contribuição dessas informações sob um olhar mais estrutural. Esse estudo é apresentado no Capítulo 8 sob o título "Estudo II - Representação de informações semânticas em conceitos formais".

Já para o aspecto "aplicabilidade da abordagem em outros *corpora* e domínios" realizamos estudos de natureza exploratória e experimental na área de categorização de documentos. A meta era analisar a efetiva contribuição das estruturas conceituais geradas a partir de nossa proposta na tarefa de classificação. Chamamos a esta fase de "Estudo III - Aplicabilidade da proposta e estudos em Língua Portuguesa" e a detalhamos no Capítulo 9. Como o título do capítulo já diz, incluímos nesse capítulo também estudos com *corpus* para Língua Portuguesa.

Cabe ressaltar que até este capítulo, os estudos mencionados foram todos realizados em *corpora* para Língua Inglesa. A principal razão para isso reside no fato de não encontramos ferramentas de anotação de papéis semânticos para Língua Portuguesa, o que era fundamental para nossa pesquisa. No entanto, realizamos alguns estudos usando *corpus* em Língua Portuguesa (Seção 9.2), focando a extração de conceitos a partir de textos e a categorização de documentos. Apesar da falta de recursos, desenvolvemos um estudo preliminar quanto ao uso de papéis semânticos em estruturas FCA a partir de informações extraídas de textos em português. Esse estudo é apresentado na Seção 9.2.4.

Cabe ressaltar também que o aspecto "avaliação da abordagem proposta" permeia tanto o Estudo II, para o qual aplicamos uma avaliação de ordem estrutural - a coesão lexical - quanto o Estudo III, no qual usamos uma avaliação de ordem funcional e aplicamos métricas usuais em categorização de documentos, tais como precisão e abrangência (*recall*) [135].

A seção seguinte descreve o pré-processamento dos textos e o estudo quantitativo dos *corpora* em Língua Inglesa Penn TreeBank Sample e SemLink 1.1, utilizados em nossa pesquisa.

6.4 Extração e análise quantitativa preliminar das informações existentes nos *corpora* Penn TreeBank Sample e SemLink 1.1

Para realizar este estudo escolhemos, o *corpus* PropBank (vide Seções 4.3 e 5.1.1.2). A escolha se baseou essencialmente no fato de esse *corpus* conter as anotações semânticas necessárias para tal estudo. Além disso, como a maioria dos etiquetadores de papéis semânticos atuais segue o formato de anotação PropBank, a aplicação de nossa abordagem a outros *corpora* se tornaria mais viável.

Apesar de facilitar a aplicabilidade de nossa proposta, o alinhamento dos rótulos numéricos providos por esta anotação ainda era um problema para o processo de agrupamento dos termos que objetiva a construção dos conceitos. Esses rótulos, ainda que idênticos, só poderiam ser unificados quando fossem associados a verbos de uma mesma classe. Nessa etapa encontramos o *corpus* SemLink 1.1 (Seção 5.1.1.3) que mapeia a VerbNet ao PropBank. Nesse *corpus*, que é uma extensão do *corpus* PropBank, foram incluídos, na anotação, a classe VerbNet do verbo e o papel semântico correspondente ao rótulo numérico PropBank. Embora esse mapeamento não seja completo, serviu para contornar o problema e viabilizou nosso estudo inicial.

Tanto o *corpus* PropBank quanto o SemLink 1.1 contêm apenas as anotações semânticas, não incluem as sentenças. Era necessário, portanto, alinhar o TreeBank-2 (Seção 5.1.1) a um desses *corpora* para extrair os termos juntamente com suas anotações. Usamos, nesse alinhamento, o *corpus* Penn TreeBank Sample (Seção 5.1.1.1) que corresponde a 9% do TreeBank-2. Optamos por este pequeno *corpus* por ser gratuito e adequado para este estudo exploratório preliminar. A Figura 6.1 mostra a sentença 0 do texto wsj_0001 e as anotações linguísticas providas pelos

corpora PennTreeBank Sample e SemLink 1.1 antes do processo de alinhamento.

Sem anotação:
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Anotação lexicossintática (Penn TreeBank Sample):
((S (NP-SBJ (NP (NNP Pierre) (NNP Vinken)) (, ,) (ADJP (NP (CD 61) (NNS years)) (JJ old)) (, ,)) (VP (MD will) (VP (VB ***join****) (NP (DT the) (NN board)) (PP-CLR (IN as) (NP (DT a) (JJ nonexecutive) (NN director))) (NP-TMP (NNP Nov.) (CD 29)))) (. .)))*

Anotação semântica (SemLink):
wsj/00/wsj_0001.mrg 0 8 auto join.01;VN=22.1-2 vf–a 0:2-ARG0[Agent] 7:0-ARGM-MOD 8:0-rel 9:1-ARG1[Patient1] 11:1-ARGM-PRD 15:1-ARGM-TMP

Figura 6.1 – Sentença 0 do texto wsj_0001 e suas anotações linguísticas (providas pelos *corpora* Penn TreeBank Sample e SemLink 1.1).

Cabe lembrar que, para cada verbo que foi anotado semanticamente em uma sentença, há uma linha no SemLink descrevendo tal anotação (Seções 5.1.1.2 e 5.1.1.3). Portanto, ao alinharmos os *corpora*, especificamos a que verbo se referem as anotações. Desta forma, conseguimos relacionar o verbo a seus argumentos e a suas informações semânticas. Na Figura 6.2, que ilustra o alinhamento dos *corpora* para a Sentença 0 do texto wsj_0001, a anotação ***{ARG0[Agent];VERB[join]*** *(NP-SBJ (NP (NNP Pierre) (NNP Vinken)) (, ,) (ADJP (NP (CD 61) (NNS years)) (JJ old)) (, ,))****}*** indica que esse segmento é um argumento do verbo *to join,* marcado com a etiqueta numérica PropBank ARG0 a qual corresponde ao papel VerbNet Agent.

((S ***{ARG0[Agent];VERB[join]*** *(NP-SBJ (NP (NNP Pierre) (NNP Vinken)) (, ,) (ADJP (NP (CD 61) (NNS years)) (JJ old)) (, ,))****}*** *(VP* ***{ARGM-MOD;VERB[join]*** *(MD will)****}*** *(VP* ***{join.01;VN=22.1-2*** *(VB join)****}*** ***{ARG1[Patient1];VERB[join]*** *(NP (DT the) (NN board))****}*** ***{ARGM-PRD;VERB[join]*** *(PP-CLR (IN as) (NP (DT a) (JJ nonexecutive) (NN director)))****}*** ***{ARGM-TMP;VERB[join]*** *(NP-TMP (NNP Nov.) (CD 29))))****}*** *(. .)))*

Figura 6.2 – Sentença 0 do texto wsj_0001 e suas anotações linguísticas, após o alinhamento dos *corpora* Penn TreeBank Sample e SemLink 1.1.

Após o alinhamento, processamos os textos a fim de extrair deles os sintagmas nominais, papéis semânticos e classes de verbos relevantes para nossa pesquisa. A partir desse pré-processamento, que é detalhado nas Seções 6.4.1 e 6.4.2, realizamos uma análise quantitativa dos *corpora* quanto aos elementos extraídos. Os resultados dessa análise são comentados na Seção 6.4.3 e são usados como critério de decisão para a escolha de sementes, de papéis semânticos e de classes de verbos usados no Estudo II.

Cabe comentar ainda que implementamos em linguagem Java os programas que realizaram tanto o alinhamento dos *corpora* quanto o pré-processamento das sentenças e a análise quantitativa.

6.4.1 Pré-processamento dos *corpora*

Após o alinhamento dos *corpora*, realizamos o pré-processamento dos textos cujo principal objetivo, a exemplo de outros trabalhos como [41, 89], era extrair, das sentenças desses textos, os verbos e seus argumentos, bem como suas anotações semânticas. Durante esse processo, consideramos apenas verbos que possuíam classe VerbNet definida. Optamos por este filtro, primeiramente, porque tais classes foram associadas apenas a verbos que receberam anotação

PropBank e, consequentemente, também possuíam argumentos etiquetados com papéis semânticos VerbNet. Outra razão é que poderíamos estudar a contribuição das classes dos verbos na construção dos conceitos formais e, ainda, comparar essa abordagem com outras, tais como as que consideram apenas os verbos, por exemplo (esse estudo é apresentado no Capítulo 8).

Para os verbos assim selecionados extraímos, dos argumentos anotados, seus sintagmas nominais e papéis semânticos associados. Decidimos trabalhar com sintagmas nominais, pois eles são considerados bons candidatos a conceitos em aprendizagem de estruturas ontológicas a partir de textos [116]. Pela mesma razão, consideramos apenas os sintagmas nominais cujos núcleos são substantivos. Outro motivo para esta escolha é que, em tarefas de classificação de texto, como a que descrevemos no Capítulo 9, o uso de n-gramas tem contribuído para a melhora dos resultados [32]. Como sintagmas nominais também são uma espécie de n-gramas, acreditamos que seu uso seja indicado para nossa pesquisa.

Com o propósito de extrair os verbos e seus argumentos, começamos o pré-processamento (Figura 6.3) identificando e normalizando morfologicamente os terminais (folhas) das árvores sintáticas das sentenças dos textos. Aplicamos o *stemmer* do pacote NLTK (Seção 5.3.2) e o lematizador Treetagger (Seção 5.3.1) aos terminais cujas etiquetas POS se referiam a verbos, substantivos comuns e adjetivos. Posteriormente, usamos os lemas desses termos para construir os conceitos formais; e usamos suas bases (*stem*s) para definir um conjunto de sementes. Os termos, etiquetados como substantivos, cujas bases eram mais frequentes, foram usados como sementes no processo de construção das estruturas conceituais (o processo de definição das sementes é detalhado na Seção 6.4.3).

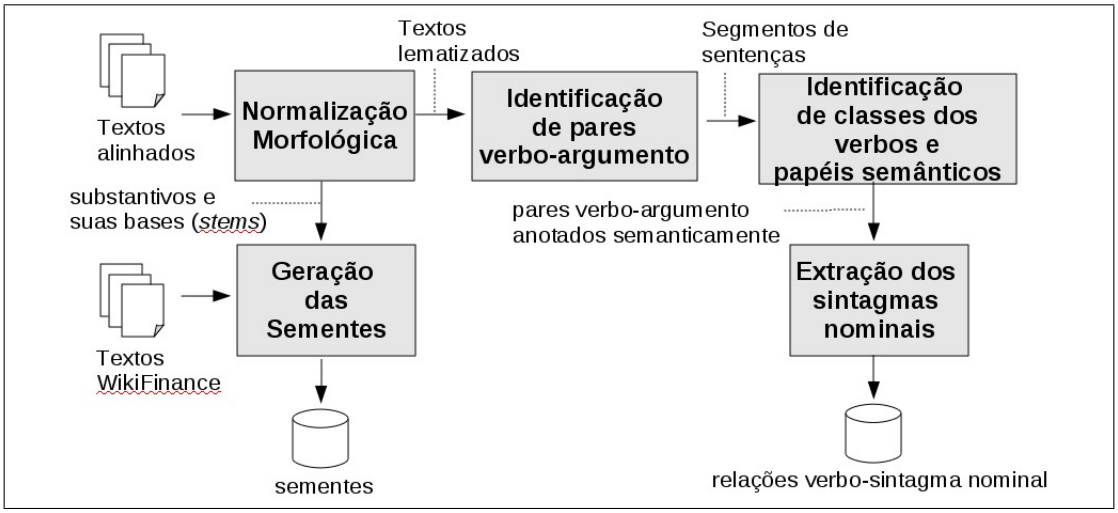


Figura 6.3 – Pré-processamento dos textos alinhados.

Terminado o processo de normalização, iniciamos, então, o processo de extração propriamente dito. A extração dos verbos e suas anotações semânticas foi simples, as etiquetas facilitaram a identificação das informações. Na sentença 4 do texto wsj_0003, exemplificada na Figura 6.6, o verbo (identificado pela etiqueta VBD) é *said*, sua classe VerbNet (marcada pela etiqueta VN) é a *37.7* e seu *frameset* corresponde ao *say.01* (informação posicionada antes da etiqueta VN). Cabe mencionar que usamos os lemas dos verbos ao definir os conceitos formais.

Já no caso dos argumentos desses verbos, tivemos que definir algumas heurísticas para a extração dos sintagmas nominais, embora eles estivessem marcados com etiquetas NP. Isso aconteceu porque, ao analisarmos as anotações semânticas atribuídas às sentenças, percebemos que os argumentos marcados com papéis temáticos não correspondiam apenas a sintagmas nominais, mas também a segmentos de sentenças. Nesses segmentos encontramos inúmeras construções gramaticais, as quais aumentaram consideravelmente a complexidade de imple-

mentação do *parser* responsável por esse processo de extração. Encontramos segmentos curtos, de tratamento computacional mais simples, delimitados por sintagmas constituídos de apenas um determinante e um substantivo comum, mas também localizamos segmentos mais longos e complexos, contendo orações e sintagmas aninhados.

Visto que a implementação de um *parser* "completo" (no sentido linguístico) com tal finalidade, ainda que relevante, não era o foco principal de nossa pesquisa, optamos por simplificar tal implementação, tratando um subconjunto dessas construções. O conjunto de heurísticas definido para tratar tais construções é o assunto da próxima seção.

6.4.2 Heurísticas para extração de sintagmas nominais

Como mencionado, tivemos que simplificar o processo de extração dos sintagmas nominais. Procuramos primar, então, pela qualidade e não pela quantidade, utilizando, na implementação desse *parser,* as heurísticas especificadas a seguir:

- desconsideramos, como já mencionado anteriormente, os verbos cuja classe VerbNet não é definida (VN=None). Logo, não processamos os argumentos desses verbos. Esse é o caso do verbo *to use* da sentença 0 do texto wsj_0003 exemplificada na Figura 6.4. Argumentos, no entanto, que são orações, ou seja, que também contêm verbos, poderão ser analisados posteriormente pelo *parser* desde que estejam anotados. Esse é o caso do argumento 4 que inclui o verbo *to make*, cujo processamento é mostrado na Figura 6.5.

- não processamos argumentos com etiquetas -NONE-, as quais marcam um elemento nulo para o qual nenhuma etiqueta POS válida foi atribuída. Esse é o caso dos argumentos 1 dos verbos *to make* e *to say* apresentados, respectivamente, na Figura 6.5 e na Figura 6.8.

- tratamos separadamente o processamento de substantivos comuns e de substantivos próprios. Durante o processamento de um sintagma nominal que contenha ambos os tipos de substantivos, como no caso do argumento 2 do verbo *to make*, apresentado na Figura 6.5, os substantivos próprios são ignorados. No caso da presença de preposições, somente a preposição "*of*" foi considerada. Quando outras preposições foram detectadas, o processamento do sintagma foi abreviado, não sendo analisado o trecho a partir da preposição. Os sintagmas nominais resultantes desse processo são, ainda, decompostos em termos aninhados. No caso de sintagmas contendo a preposição "*of*", somente é decomposto o n-grama anterior à preposição. Os substantivos próprios são extraídos somente quando o sintagma não tem substantivos comuns (Figura 6.8).

- descartamos argumentos que correspondam a segmentos contendo outras orações. As orações são identificadas pela presença de verbos. É importante frisar que os argumentos são desconsiderados somente quando não há qualquer pontuação anterior à ocorrência do primeiro verbo. Encontramos argumentos contendo orações basicamente em duas situações: quando o verbo presente no segmento não possui nenhum tipo de anotação semântica, como *to be*; ou quando a sentença possui anotações semânticas aninhadas, ou seja, o verbo do segmento também tem argumentos anotados semanticamente. Esse último caso já foi mencionado em item anterior (caso do verbo *to make* que aparece no argumento 4 do verbo *to use* e é processado posteriormente pelo *parser*). A primeira das duas situações, no entanto, é exemplificada na Figura 6.6. O *parser,* ao processar a sentença 4 do texto wsj_0003, despreza o *argumento 2* do verbo *to say*, pois este argumento possui um verbo (*to be*) sem anotação semântica. Sendo extraídos, portanto, somente os sintagmas nominais do argumento 1 do verbo *to say*.

Sentença completa e sem anotação: *A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago, researchers reported.*

Trecho referente ao verbo ***to use***: *A form of asbestos once used to make Kent cigarette filters ...*

- Verbo: *{**use.01;VN=None** (VBN used)}*
  - Argumento 1: *{**ARGM-TMP** (ADVP-TMP (RB once)) }*
  - Argumento 2: {**ARG1** *(NP (NP (DT A) (NN form)) (PP (IN of) (NP (NN asbestos))))*
  - Argumento 3: *{**ARG1** (-NONE- *) }*
  - Argumento 4: *{**ARG2-PNC** (VP (VB make) (NP (NNP Kent) (NN cigarette) (NNS filters))) }*
- Sintagmas nominais extraídos: {}

Figura 6.4 – Anotações da sentença 0 do texto wsj_0003 para o verbo *to use* .

Trecho referente ao verbo ***to make***: *... make Kent cigarette filters ...*

- Verbo: *{**make.01;VN=26.1-1** (VB make)}*
  - Argumento 1: *{**ARG0[Agent]** (-NONE- *) }*
  - Argumento 2: *{**ARG1[Product]** (NP (NNP Kent) (NN cigarette) (NNS filters))}*
- Sintagmas nominais extraídos: {*cigarette_filters, filters*}

Figura 6.5 – Anotações da sentença 0 do texto wsj_0003 para o verbo *to make* .

- observamos os sinais de pontuação (vírgula, parênteses, ...) existentes nos segmentos definidos como argumentos. Se o segmento em questão não for constituído apenas de substantivos próprios, seu processamento é abreviado. O *parser* analisa o segmento até a ocorrência de um sinal de pontuação. O texto restante do argumento, posterior a essa pontuação, é descartado. Esse é o caso do argumento 2 do verbo *to propose* da sentença 13 do texto wsj_0013, apresentado na Figura 6.7.

- tratamos enumerações de substantivos comuns, usuais em sujeitos compostos, desde que mediadas por conjunções *and* e *or*. As enumerações separadas por sinais de pontuação são processadas apenas até a ocorrência do primeiro sinal, como descrito em item anterior.

- processamos construções que possuam um aposto explicativo, desde que o segmento correspondente a essa construção contenha um sintagma nominal formado apenas por substantivos próprios e um aposto justapostos. Esse é o caso do argumento 2 do verbo *to say* da sentença 27 do texto wsj_0003 (Figura 6.8). O foco de nosso processamento não está nos substantivos próprios. No entanto, para alguns estudos, precisamos da relação instância-classe, para a qual tais substantivos são relevantes. Escolhemos, por simplicidade de processamento, esse tipo de construção para extrair essas relações. Do segmento referente ao aposto, extraímos as classes, ou seja, os sintagmas contendo substantivos comuns; e usamos os sintagmas formados por substantivos próprios como suas instâncias.

Sentença sem anotação: *A Lorillard spokewoman said, "This is an old story."*

- Verbo: *{**say.01;VN=37.7** (VBD said)}*
  - Argumento 1: *{**ARG0[Agent]** (NP-SBJ (DT A) (NNP Lorillard) (NN spokewoman))}*
  - Argumento 2: {**ARG1[Topic]** (S (NP-SBJ (DT This)) (VP (**VBZ is**) (NP-PRD (DT an) (JJ old) (NN story))))}
- Sintagmas nominais extraídos para o papel ARG0[Agent]: *{spokewoman}*

Figura 6.6 – Sentença 4 do texto wsj_0003 e suas anotações linguísticas.

Sentença completa e sem anotação: *The fact that New England proposed lower rate increases – 4.8% over seven years against around 5.5% boosts proposed by the other two outside bidders – complicated negotiations with state officials, Mr. Ross asserted.*

Trecho referente ao verbo **to propose**: *The fact that New England proposed lower rate increases – 4.8% over seven years against around 5.5% boosts proposed by the other two outside bidders – ...*

Trecho descartado: *... – 4.8% over seven years against around 5.5% boosts proposed by the other two outside bidders – ...*

- Verbo: *{**propose.01;VN=37.7** (VBD proposed)}*
  - Argumento 1: *{**ARG0[Agent]** (NP-SBJ (NNP New) (NNP England))}*
  - Argumento 2: *{**ARG1[Topic]** (NP (NP (JJR lower) (NN rate) (NNS increases)))}*
- Sintagmas nominais e termos aninhados extraídos para o papel ARG0[Agent]: {*New_England*}
- Sintagmas nominais e termos aninhados extraídos para o papel ARG1[Topic]: {*increase, rate_increase, low_rate_increase*}

Figura 6.7 – Anotações da sentença 13 do texto wsj_0013 para o verbo *to propose*.

- excluímos, dos sintagmas nominais extraídos, substantivos relacionados a tempo, como *week*, *day*, *month*, *year, today* etc. Também excluímos outros elementos menos significativos para a formação de conceitos, como caracteres especiais, numerais, artigos e pronomes. Se, após a eliminação desses elementos, o sintagma nominal restante é inválido, principalmente devido à ausência de substantivos, o argumento também é descartado. Cabe mencionar que termos referentes a tempo são muito frequentes em textos jornalísticos, no entanto sua relevância tende a ser baixa para nossa pesquisa, visto que não tratamos a ordem cronológica em que ocorrem os fatos expressos nos textos. Como também não tratamos numerais, quando esses termos se referem a quantidades, eles igualmente se tornam irrelevantes.

- também não consideramos os sintagmas nominais presentes em argumentos anotados com etiquetas com final TMP, como NP-TMP ou PP-TMP, pois, igualmente, são relativas a

tempo. Essas etiquetas marcam segmentos da sentença que determinam quando, quantas vezes, ou por quanto tempo algo ocorreu.

Sentença completa e sem anotação: *"There's no question that some of those workers and managers contracted asbestos-related diseases," said Darrell Phillips, vice president of human resources for Hollingsworth & Vose.*

Trecho referente ao verbo **to say**: *... said Darrell Phillips, vice president of human resources for Hollingsworth & Vose.*

- Verbo: *{**say.01;VN=37.7** (VBD said)}*
  - Argumento 1: *{**ARG1[Topic]** (-NONE- *) }*
  - Argumento 2: {**ARG0[Agent]** *(NP-SBJ (NP (NNP Darrell) (NNP Phillips)) (, ,) (NP (NP (NN vice) (NN president)) (PP (IN of) (NP (JJ human) (NNS resources))) (PP (IN for) (NP (NNP Hollingsworth) (CC &) (NNP Vose)))))* }
- Sintagmas nominais e termos aninhados extraídos para o papel ARG0[Agent]: {*Darrell_Phillips, vice_president_of_human_resource, vice_president, president*}

Figura 6.8 – Anotações da sentença 27 do texto wsj_0003 para o verbo *to say* .

O processo de extração de informações, conforme descrito nesta seção, resultou em 11.076 relações do tipo verbo-argumento. Visto que estudamos também as relações entre os papéis semânticos, formamos tuplas a partir dessas relações. Cada tupla é constituída de um verbo, sua classe VerbNet e dois de seus argumentos (sintagmas nominais constituídos de substantivos comuns) existentes em uma mesma sentença. Para cada argumento, associamos os papéis semânticos VerbNet e PropBank correspondentes. Anexamos ainda, aos argumentos, suas instâncias (sintagmas constituídos de substantivos próprios), quando elas existem. Também documentamos, para fins de análise, informações quanto às sentenças e textos a partir dos quais tais tuplas foram extraídas. A Tabela C.6 do Apêndice C apresenta exemplos dessas tuplas.

Após esse processo de extração, analisamos as relações expressas pelas tuplas e decidimos quais das informações obtidas poderiam ser interessantes para nossa pesquisa. Essa análise é o assunto da próxima seção.

6.4.3 Informações extraídas dos *corpora*

O conjunto de sentenças anotadas que tínhamos para este estudo inicial era relativamente pequeno. Logo, fizemos um levantamento a fim de determinar quais das informações extraídas poderiam prover resultados quantitativamente mais significativos.

Durante o processamento das anotações morfossintáticas das 3.914 sentenças pertencentes aos 199 textos do Penn TreeBank Sample, encontramos 100.673 *tokens*. Desse total, apenas 94.081 *tokens* estavam associados a etiquetas POS válidas. Analisamos, então, a distribuição desses *tokens* em categorias lexicais, contabilizando suas etiquetas POS (Tabela C.1 do Apêndice C). Nosso interesse, no entanto, estava nas quantidades relativas a substantivos, adjetivos e verbos, os quais tinham maior relevância para nosso estudo. A Tabela 6.1 apresenta a frequência dessas categorias no Penn TreeBank Sample.

Em seguida, precisavamos determinar quais dos substantivos comuns poderiam ser usados como sementes na construção das estruturas conceituais. Imaginamos que os mais frequentes

Tabela 6.1 – Quantidade de substantivos, adjetivos e verbos existentes no Penn TreeBank Sample

Classe gramatical (etiqueta POS)	#
Substantivos Comuns (NN e NNS)	19.213
Verbos (VB, VBD, VBZ, VBG, VBN, VBP)	12.637
Substantivos Próprios (NNP e NNPS)	9.653
Adjetivos (JJ)	5.834

poderiam conter uma gama maior de relacionamentos tanto com verbos quanto com outros substantivos e, portanto, poderiam gerar estruturas FCA semanticamente mais relevantes.

Para definir as sementes, usamos os *stems* dos termos etiquetados como substantivos comuns. Analisamos aqueles termos cujos *stems* eram mais frequentes, ou seja, que possuíam mais de 50 ocorrências[1]. Escolhemos, entre os 57 termos assim definidos, os mais relevantes para o domínio de Finanças. Para determinar tal relevância, usamos o *corpus* WikiFinance (Seção 5.1.2.1) como referência. Com esse fim, construímos uma lista com os termos desse *corpus* ordenados pela frequência de seus *stems*. Comparamos, então, os termos do WikiFinance aos 57 do Penn TreeBank Sample e selecionamos os 10 mais frequentes[2] em ambos os *corpora*. Os termos escolhidos e suas frequências são apresentados na Tabela 6.2. Cabe mencionar que os 10 termos escolhidos possuem frequência superior a 400 no WikiFinance. Cabe destacar também que desconsideramos, nessa seleção, termos referentes a tempo, como *year* e *month*.

Tabela 6.2 – Sementes

Termo	#	**Termo**	#
company	324	price	162
market	248	rate	115
share	231	bank	108
stock	207	investor	105
trade	204	fund	96

Precisávamos ainda determinar quais informações semânticas relativas aos verbos poderiam ser mais significativas para nosso estudo. Analisamos, então, as anotações providas pelo *corpus* SemLink 1.1. Das 112.917 linhas de anotação semântica contidas nesse *corpus*, 9.353 correspondiam às sentenças do Penn TreeBank Sample.

Iniciamos analisando as classes dos verbos. Das 12.637 instâncias de verbos, 8.492 (~67%) possuíam anotação de classe VerbNet. Encontramos 245 classes diferentes: 178 classes principais e 67 subclasses (Figura C.1 do Apêndice C). Os verbos pertencentes a subclasses, no entanto, eram menos frequentes. Verificamos que cerca de 80% das 11.076 relações extraídas continham verbos anotados com classes principais da VerbNet.

Para visualizar a expressividade de cada classe, contabilizamos as instâncias de seus verbos (Tabela C.2 do Apêndice C). As 10 classes VerbNet mais frequentes são apresentadas no gráfico da Figura 6.9. A classe mais frequente é a 37.7. Encontramos nos textos 18 verbos dessa classe, tais como: *to say*, *to propose*, *to announce*, *to suggest*, *to claim*, *to disclose* e *to report*. A alta frequência da classe 37.7 já era esperada, visto que ela é formada por verbos muito usados em textos jornalísticos.

A segunda classe mais frequente é a 45.4. Para ela encontramos mais tipos de verbos: 73. Fazem parte dessa classe verbos como *to close*, *to improve*, *to operate*, *to increase* e *to*

[1] O valor relativo à frequência dos termos analisados (mais de 50 para o *corpus* Penn TreeBank Sample) foi arbitrária.

[2] Optamos por usar apenas 10 termos como sementes, igualmente de forma arbitrária.

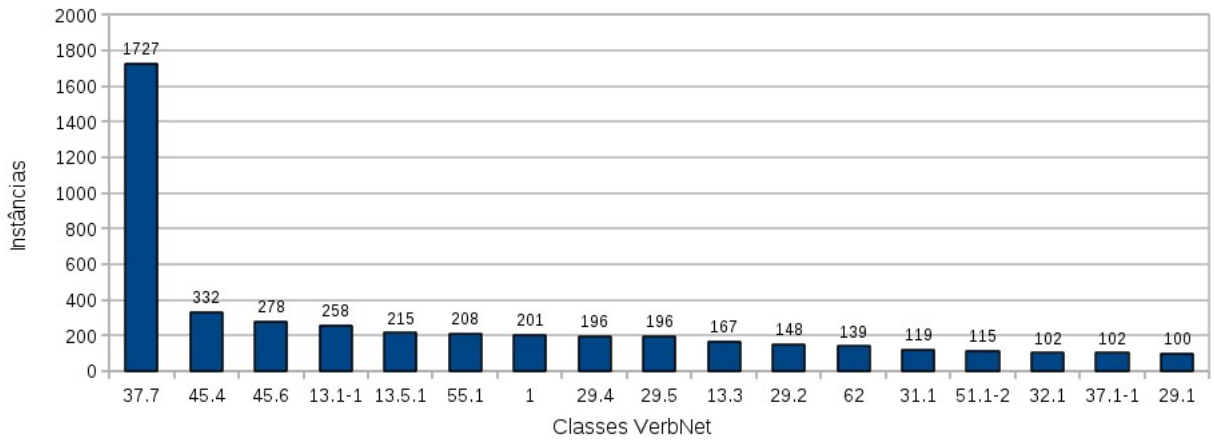


Figura 6.9 – As 20 classes VerbNet mais frequentes nos *corpora* analisados.

open. Analisando os verbos dessa classe, observamos que vários descrevem ações comuns ao domínio de Finanças. Em razão, principalmente, da frequência, escolhemos as classes 37.7 e 45.4 para estudar mais detalhadamente. Esse estudo é apresentado no Capítulo 7. A Tabela C.3 apresenta a lista completa dos verbos para as 5 classes mais frequentes no *corpus* Penn TreeBank Sample.

Em seguida, analisamos os papéis semânticos associados aos argumentos dos verbos. Ao processar o *corpus* SemLink 1.1 Sample[3], identificamos 22 papéis semânticos. O significado de cada uma dessas etiquetas semânticas foi descrito na Tabela C.4 do Apêndice C. Observamos ainda a ocorrência desses papéis nos *corpora*. O gráfico da Figura 6.10 mostra os papéis semânticos identificados e suas respectivas frequências[4].

Dos 9 papéis mais frequentes, escolhemos 8 para analisar: Agent, Theme, Patient, Topic, Experiencer, Recipient, Cause e Product. Excluímos Predicate por julgá-lo pouco informativo. Observamos, então, a distribuição desses papéis quanto às classes dos verbos e às instâncias dos argumentos (Tabela 6.3). Com base nessa distribuição, concentramos nosso estudo nos papéis Agent, Theme, Patient e Topic. No entanto, há momentos em nossa investigação, em que todos os papéis semânticos são considerados, tal como nos estudos relativos às classes VerbNet dos verbos.

Os papéis Agent e Theme são muito frequentes, aparecem associados a quase um terço dos argumentos. Além disso, os verbos que definem tais papéis pertencem a mais de 40% das classes VerbNet identificadas. Essas distribuições poderiam permitir um estudo mais abrangente pelo fato de os papéis permearem várias classes, e também um estudo provavelmente mais conclusivo devido à frequência desses papéis em argumentos. Já os papéis Patient e Topic são menos frequentes em argumentos, mas estão concentrados em classes que escolhemos estudar. Patient, por exemplo, aparece associado a argumentos de vários verbos da classe 45.4 e Topic aparece como papel recorrente dos argumentos de verbos da classe 37.7.

[3]Cabe lembrar, SemLink 1.1 Sample foi o nome que atribuímos ao subconjunto do *corpus* SemLink 1.1 cujas anotações correspondem às sentenças do Penn TreeBank Sample.

[4]As variantes Theme1 e Theme2, bem como Patient1 e Patient2, foram contabilizadas, respectivamente, com os papéis a partir dos quais são derivadas: Theme e Patient. Conforme a documentação VerbNet, disponível em http://verbs.colorado.edu/~mpalmer/projects/verbnet.html#thetaroles, essas variantes de papéis são usadas por algumas classes VerbNet, para as quais os verbos possuem dois argumentos que, por falta de uma clara distinção entre os mesmos, recebem o mesmo papel. Considerando as variantes, são totalizados 28 papéis. A Tabela C.5 mostra esses papéis e suas frequências, no Apêndice C.

Tabela 6.3 – Percentuais de distribuição dos papéis semânticos nas classes de verbos e nos argumentos.

Papel Semântico	Classes (%)	Instâncias de argumentos (%)
Agent	51	28
Theme	44	30
Patient	12	6
Topic	9	14
Experiencer	8	2
Recipient	7	2
Cause	5	2
Product	4	2

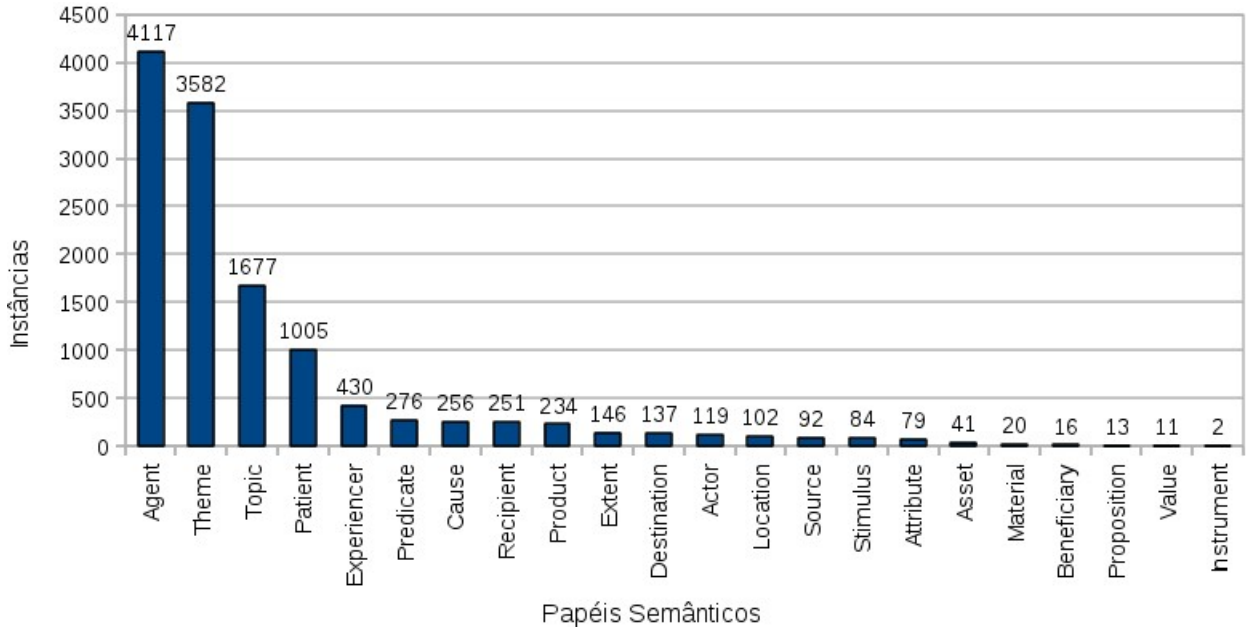


Figura 6.10 – Frequência das instâncias de papéis semânticos encontrada nos *corpora* Penn TreeBank Sample e SemLink 1.1 Sample.

Cabe mencionar que não consideramos em nosso estudo os papéis ARGM-MOD (modificador de modo) e ARGM-TMP (modificador de tempo), pois entendemos que o modo é pouco relevante para a construção de conceitos e, como já mencionamos, não tratamos termos relacionados a tempo. Os argumentos associados a esses papéis também foram descartados.

6.5 Considerações sobre este capítulo

Analisando os dados quantitativos apresentados nesse capítulo, acreditamos que a natureza do *corpus* tenha uma grande influência na frequência de determinados papéis. Como o *corpus* Penn TreeBank Sample é de cunho jornalístico, os papéis Agent, Theme e Topic são naturalmente os mais recorrentes. Observamos que 44% das instâncias de argumentos presentes em relações verbo-argumento estão associadas aos papéis Theme e Topic.

Como os papéis Theme e Topic, de acordo com suas definições na Tabela C.4 do Apêndice C, são mais genéricos, a natureza do *corpus,* de uma certa forma, aumenta quantitativamente a importância desses papéis e, portanto, "impede" que relações semânticas de domínio mais interessantes se destaquem. Embora as relações do tipo Agent-Theme e Agent-Topic nos tragam informações interessantes na medida que estabelecem ligações entre pessoas ou entidades a

termos do domínio, pouco se pode dizer sobre os argumentos anotados como Theme e Topic nessas relações.

Sob esse ponto de vista, torna-se difícil determinar conclusivamente que relações semânticas são mais importantes, pelo menos quantitativamente, para o domínio de Finanças. De maneira geral, essa análise quantitativa não nos permite responder, de forma contundente, à maioria das questões referentes à identificação das relações não taxonômicas que foram apresentadas na Seção 6.2. Acreditamos que uma análise quantitativa seria mais expressiva no caso de um *corpus* de natureza conceitual.

Acreditamos também que no caso específico dos *corpora* utilizados, a seleção das relações transversais a serem usadas na construção de estruturas conceituais teria que ser baseada essencialmente no tipo dos papéis envolvidos. No contexto desses *corpora*, os papéis Theme e Topic, por exemplo, são pouco expressivos, pois são atribuídos indiscriminadamente a diferentes elementos do domínio. Outra alternativa seria solicitar a ajuda de um especialista no domínio para estabelecer as relações entre papéis mais relevantes.

Apesar de não serem conclusivos, os resultados apresentados nesse capítulo foram úteis para delinear o Estudo I. O Estudo I é abordado no capítulo a seguir.

7. ESTUDO I - ANÁLISE DE ESTRUTURAS CONCEITUAIS RCA E DE CLASSES DE VERBOS

Este capítulo descreve a investigação realizada para determinar a viabilidade de uso da extensão RCA no que se refere à inclusão de papéis semânticos em conceitos formais. Este estudo é apresentado na Seção 7.1. Apresentamos ainda, nesse capítulo, o estudo das classes de verbos 37.7 e 45.4, as quais foram consideradas, pela análise realizada no capítulo anterior, mais significativas do ponto de vista quantitativo. Estudamos os verbos e os papéis semânticos associados a essas classes, respectivamente, nas Seções 7.2 e 7.3.

7.1 Análise de estruturas conceituais RCA

Para analisarmos a viabilidade de uso da extensão RCA, criamos um pequeno exemplo focado no papel semântico Agent. Com o propósito de delimitar o contexto semântico do exemplo, usamos o termo mais frequente do *corpus* Penn TreeBank Sample - *company* - como semente inicial. Aplicamos a essa semente, então, o "operador mais" (comentado na Seção 3.6.3). Para reduzir a complexidade da estrutura FCA resultante e centralizar a análise no papel Agent, estabelecemos restrições ao "operador mais". O operador só recuperaria tuplas do tipo *verbo-argumento1-argumento2,* em que: um dos argumentos fosse a semente, esse argumento estivesse anotado com o papel Agent, e sua relação com o outro argumento ocorresse ao menos 3 vezes no *corpus*. Cabe mencionar que, para simplificar nossa análise, consideramos, no exemplo, apenas unigramas como argumentos.

A partir das tuplas selecionadas pelas aplicações sucessivas do "operador mais", o qual utilizava os argumentos recuperados como novas sementes, criamos dois contextos formais. Um dos contextos, denominado *company*, relacionava os argumentos (objetos) aos verbos correspondentes (atributos). Esse contexto foi definido a partir dos conjuntos G = *{share, shareholder, market, pence, company}* e M = *{receive, place, buy, rise, offer, redeem, pay}*.

Construímos também o contexto relacional "is_agent_of" que conecta os objetos em G conforme tal relação. Cabe destacar que a relação "is_agent_of" não é simétrica e nem reflexiva. Ela foi definida formalmente como $c_1 R^{is\_agent\_of}[||\geq 1||, ||\geq 1||;]c_2$, onde os conceitos $c_1, c_2 \in \mathscr{B}(G, M, I)$. Os contextos assim definidos e a estrutura RCA gerada a partir deles são apresentados, respectivamente, nas Figuras 7.1 e 7.3a. A estrutura RCA foi gerada com a ferramenta ERCA (Seção 5.3.5).

```
FormalContext company
|           |receive|place|buy|rise|offer|redeem|pay|
|share      |   x   |  x  | x | x  |  x  |  x   | x |
|shareholder|   x   |     | x |    |     |      |   |
|market     |       |  x  |   |    |     |      |   |
|pence      |       |     |   | x  |     |      |   |
|company    |       |     |   |    |  x  |  x   | x |

RelationalContext is_agent_of
source share
target share
scaling com.googlecode.erca.framework.algo.scaling.Wide
|           | share | shareholder | market| pence | company|
|share      |       |             |       |       |        |
|shareholder|    x  |             |       |       |        |
|market     |       |             |       |       |        |
|pence      |       |             |       |       |        |
|company    |    x  |             |       |       |        |
```

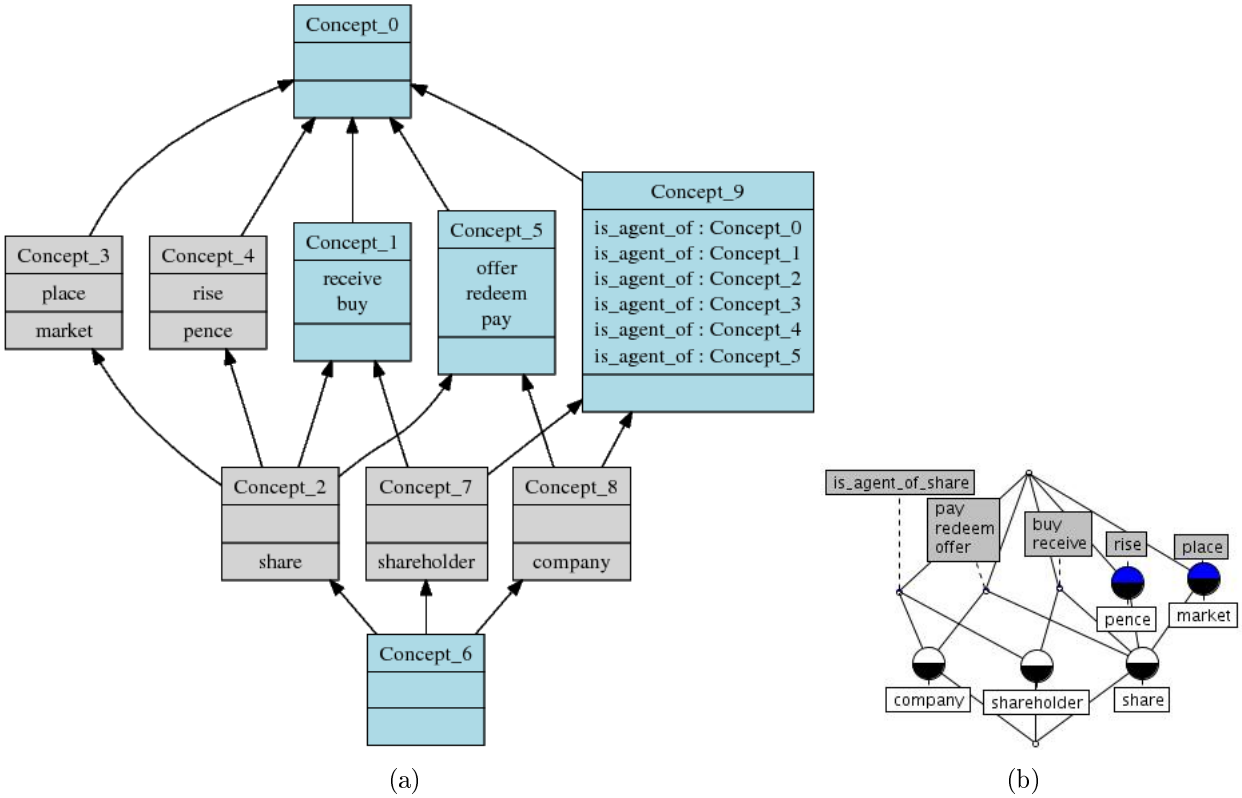
Figura 7.1 – Contextos para geração da estrutura RCA definidos a partir da semente *company*.

Em seguida, para que pudéssemos comparar a estrutura FCA com sua extensão RCA, criamos um novo contexto formal. Este novo contexto (Figura 7.2) foi definido a partir do "contexto formal *company*". Ele possui um atributo a mais: a relação "is_agent_of_share". A estrutura FCA correspondente a esse contexto é apresentada na Figura 7.3b e foi construída com a ferramenta Concept Expert (Seção 5.3.5).

	receive	place	buy	rise	offer	redeem	pay	is_agent_of_share
share	X	X	X	X	X	X	X	
shareholder	X		X					X
market		X						
pence				X				
company					X	X	X	X

Figura 7.2 – Contexto formal da estrutura FCA definido a partir da semente *company*.

Analisando as estruturas, percebemos que foi gerado coincidentemente o mesmo número de conceitos (10). Como já esperavámos, exceto pelo Concept_9, todos os demais conceitos da estrutura RCA têm um conceito correspondente na estrutura FCA. A grande diferença, que inclusive já foi comentada na Seção 3.4, é que, no RCA, *is_agent_of* é uma relação entre conceitos e, no FCA, *is_agent_of_share* expressa uma relação entre *share* e os demais objetos. Como, nesse exemplo, existe Agent somente para *share*, apenas um atributo com esse fim foi criado. No entanto, se mais agentes existissem para os demais objetos, cada uma essas relações teria que ser especificada como um novo atributo no contexto formal da estrutura FCA. Cabe mencionar que, como o papel semântico Agent foi especificado com um atributo, formalmente, não há uma relação dos demais objetos com o objeto *share*. *Share*, nesse caso, é simplesmente parte do rótulo do atributo Agent.



(a) (b)

Figura 7.3 – Estruturas RCA e FCA para a semente *company*.

Já no caso RCA, para incluir mais agentes, bastaria modificar o contexto relacional *is_agent_of* assinalando tais relações. Apesar dessa facilidade e da aparente conveniência de estabelecer relações entre conceitos, a estrutura estabeleceu generalizações da relação *is_agent_of*

que nos pareceram propícias a erros de interpretação. Considerando que o Concept_9 é formado pelos objetos *company* e *shareholder*, e o Concept_1, pelos os objetos *share* e *shareholder*, ao determinar que o Concept_9 é agente de Concept_1, não é explícito que essa relação originalmente valeria apenas para *share*. Apesar de o quantificador especificar que a relação não se aplica a todos os objetos, a estrutura RCA resultante não deixa claro, por exemplo, que *company* não desempenha o papel de agente no contexto semântico de *shareholder*.

Se os objetos presentes em um mesmo conceito fossem sinônimos, a generalização de uma relação entre objetos para uma relação entre conceitos seria oportuna. No entanto, como os atributos usados são verbos, os grupos de objetos de um conceito, na maioria das vezes, não são formados por sinônimos, apenas por elementos que pertencem a um mesmo contexto semântico. Para decidir pela aplicação do método RCA, nos parece mais adequada a utilização de uma variedade maior atributos que possam especializar ainda mais os conceitos. Isso implicaria o uso de um número maior de características e não apenas de verbos, como é o caso de nosso estudo.

Mesmo conscientes de que examinamos apenas um pequeno exemplo, acreditamos que os possíveis erros de interpretação decorrentes dessa generalização são especialmente preocupantes em caminhamentos automáticos sobre a estrutura conceitual. Visto que esse é um dos nossos objetivos, resolvemos não usar a extensão RCA em nossa pesquisa.

Nas próximas seções analisamos as classes de verbos 37.7 e 45.4.

7.2 Análise da classe VerbNet 37.7

A classe 37.7 da VerbNet, como já mencionado, é a mais frequente no *corpus* Penn TreeBank Sample. Ela é composta essencialmente por verbos de comunicação, o que justifica a sua alta frequência (1.727 instâncias), tendo em vista que se trata de um *corpus* jornalístico.

Durante o pré-processamento dos textos, encontramos 18 verbos associados a esta classe. A Tabela C.3 do Apêndice C contém esses verbos e suas respectivas frequências. No entanto, após a extração dos sintagmas nominais e descarte de argumentos, este número caiu para 12 verbos e a quantidade de instâncias foi reduzida a quase um terço, ou seja, 544.

Apesar da redução, a proporção de distribuição dos verbos em instâncias foi mantida, e foi sobre esses dados que realizamos nosso estudo. Para simplificar a análise, consideramos apenas unigramas como argumentos dessas instâncias. Iniciamos o estudo, verificando a distribuição dos argumentos em papéis semânticos.

Segundo dados da VerbNet, os verbos dessa classe costumam associar 3 tipos de papéis semânticos aos seus argumentos: Agent, Topic e Recipient. Os papéis Agent e Recipient são usados para argumentos que representam algo animado ou uma organização. Como os verbos dessa classe são de comunicação, o papel Agent geralmente indica o elemento emissor, o papel Recipient, o receptor e o papel Topic, o assunto dessa comunicação.

A Tabela 7.1 exibe os verbos e a respectiva distribuição das 544 instâncias em papéis semânticos para esta classe. Observando essas instâncias, percebe-se que o papel Agent é o mais frequente. Ele aparece associado a 464 argumentos. O segundo mais frequente é o Topic, com 77 anotações. E, por fim, o papel Recipient aparece associado a apenas 3 argumentos. Essa distribuição era esperada, pois é natural que, em um *corpus* jornalístico, existam mais declarações, ou seja, mais elementos emissores do que receptores.

O verbo *to say* é o mais frequente na classe 37.7 e no *corpus* Penn TreeBank Sample. Observando-se a estrutura FCA[1]apresentada na Figura 7.4a, pode-se perceber que o verbo *to say* está relacionado a uma variedade de sintagmas nominais maior do que os demais verbos da classe. Cerca de 84% dos sintagmas relacionados a verbos desta classe são argumentos do

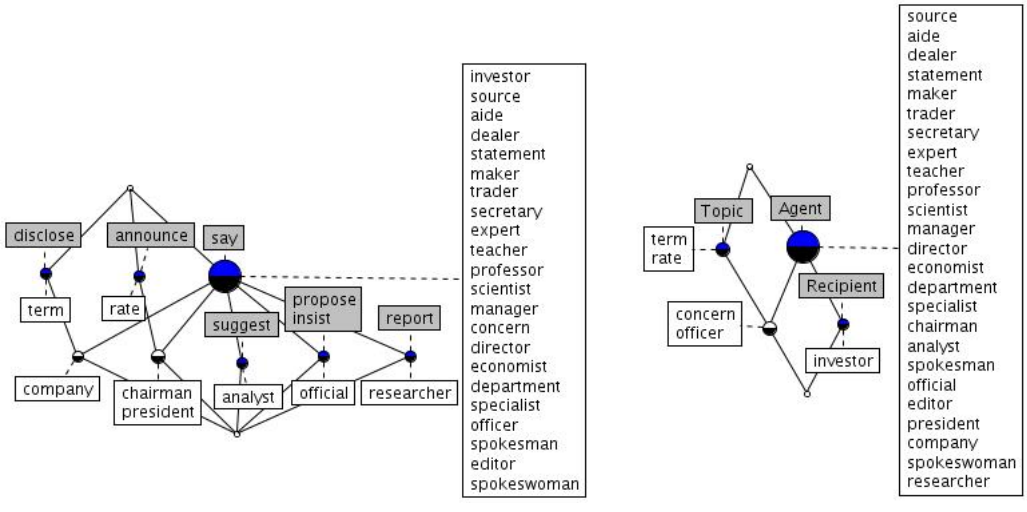
[1]Com o objetivo de gerar estruturas FCA mais "enxutas" e com sintagmas mais frequentes, as estruturas da Figura 7.4 foram geradas apenas com os sintagmas que apareciam em ao menos 3 relações verbo-argumento.

Tabela 7.1 – Distribuição dos papéis semânticos para os verbos da classe VerbNet 37.7

verbo	#Agent	#Topic	#Recipient	Total por verbo
to say	422	32	1	455
to disclose	2	15	0	17
to propose	3	14	0	17
to suggest	11	1	0	12
to announce	7	6	0	13
to claim	5	3	0	8
to report	5	0	2	7
to declare	3	2	0	5
to insist	3	0	0	3
to mention	0	3	0	3
to state	2	1	0	3
to observe	1	0	0	1
Total por papel semântico	464	77	3	-

verbo *to say*. À maioria desses sintagmas foi associado o papel Agent, conforme pode-se notar na estrutura FCA apresentada na Figura 7.4b.

Os sintagmas assim anotados correspondem, em sua maioria, a funções, cargos e profissões exercidos por pessoas (*chairman*, *director*, *professor*,...). Há aqueles sintagmas, no entanto, que expressam "denominações", mais usuais em jornais, como o caso de *source,* que é usado para indicar uma fonte de informação (exemplo: *..., U.S. sources said.* - trecho da sentença 20 do texto wsj_0093). Os agentes ainda podem ser entidades, como *company*, por exemplo. Há casos, porém, em que os agentes são entidades, mas a extração isolada da relação verbo-argumento não permite tal interpretação. Esse é o caso de *statement.* Em geral, refere-se ao comunicado de alguma empresa ou instituição (exemplo: *"..., the statement said."* - trecho da sentença 32 do texto wsj_0109).



(a) (b)

Figura 7.4 – Estruturas FCA para classe VerbNet 37.7.

De acordo com as amostras que analisamos, o papel Topic parece indicar elementos do domínio Finanças que são tópico de alguma discussão, como *rate*, por exemplo. No entanto, percebemos que nossas heurísticas para extração dos sintagmas nominais para esse papel semântico não foram muito efetivas. Verificamos que o papel Topic era constantemente atribuído a segmentos longos de sentenças. Logo, a simplificação que fizemos não foi adequada para ex-

trair informações relevantes para esse papel. Sendo, inclusive, uma das principais razões para a presença de termos pouco significativos, como *term* e *concern,* sob a etiqueta Topic.

Para o papel Recipient, no entanto, não realizamos análise dado ao pequeno número de termos associados a essa etiqueta semântica.

7.3 Análise da classe VerbNet 45.4

Como já mencionado, esta classe também é muito frequente no *corpus* Penn TreeBank Sample. Seus verbos descrevem ações comuns ao domínio de Finanças, tais como: *to open, to close, to improve, to increase,* etc.

Encontramos inicialmente 332 instâncias de 73 verbos dessa classe (Tabela C.3 do Apêndice C). Após o pré-processamento, foram descartadas 20% dessas instâncias, restando 265. Mesmo assim, a variedade de verbos ainda se manteve, embora a quantidade tenha caído para 61.

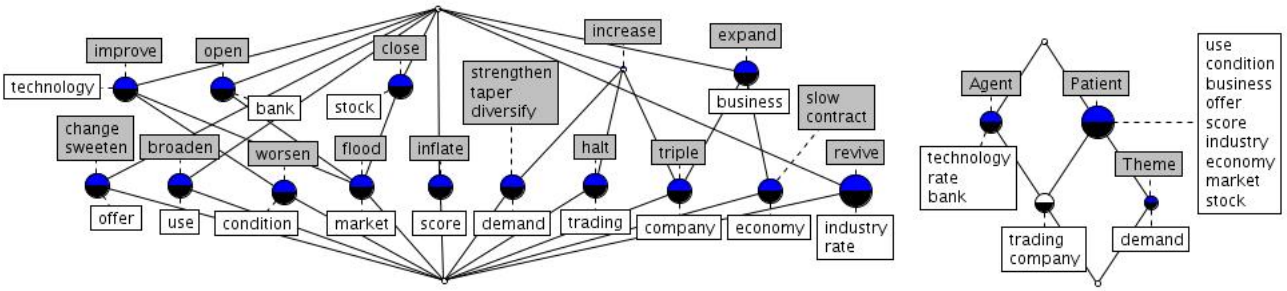
De acordo com a VerbNet, os verbos dessa classe costumam associar aos seus argumentos os papéis semânticos: Agent, Patient e Instrument. No entanto, não encontramos nas 265 instâncias analisadas, argumentos anotados como Instrument. Por outro lado, tivemos 2 argumentos anotados com o papel Theme. O papel mais frequente nesta classe foi Patient. Este foi associado a 215 argumentos. Os 48 argumentos restantes foram etiquetados como Agent. A Figura 7.5 apresenta os 61 verbos e também o tipo e a quantidade de papéis semânticos associados a seus argumentos.

to improve: Agent(6), Patient(24); *to increase*: Agent(7), Patient(18); *to expand*: Agent(1), Patient (20); *to operate*: Agent(4), Patient(16); *to grow:* Patient (18); *to close:* Agent (2), Patient (10); *to open:* Agent (2), Patient (10); *to change:* Agent (2), Patient (9); *to slow:* Agent (3), Patient (7); *to advance:* Patient (7); *to revive:* Agent (2), Patient (4); *to reopen:* Patient (4); *to weaken:* Agent (1), Patient (3); *to air:* Agent (2), Patient (1); *to blur:* Agent (2), Patient (1); *to broaden:* Patient (3); *to clear:* Agent (1), Patient (2); *to divide:* Agent (1), Patient (2); *to halt:* Agent (1), Patient (2); *to mature:* Patient (3); *to narrow:* Agent (1), Patient (2); *to sink:* Agent (1), Patient (2); *to sweeten:* Patient (3); *to tighten:* Agent (1), Patient (2); *to triple:* Agent (1), Patient (2); *to alter:* Agent (1), Patient (1); *to burn:* Agent (1), Patient (1); *to chill:* Agent (1), Patient (1); *to cool:* Patient (2); *to diminish:* Agent (2); *to double:* Agent (1), Patient (1); *to fill:* Patient (2); *to freeze:* Patient (2); *to lessen:* Agent (1), Patient (1); *to level:* Patient (2); *to mobilize:* Patient (2); *to stretch:* Patient (2); *to taper:* Theme (2); *to abate, to accelerate, to centralize, to contract, to dissolve, to diversity, to ease, to fade, to fatten, to flood, to heat, to heighten, to inflate, to lengthen, to magnify, to polarize, to rekindle, to reverse, to ripen, to soften, to strengthen, to vary, to worsen*: Patient (1).

Figura 7.5 – Verbos da classe VerbNet 45.4 e distribuição de seus argumentos em papéis semânticos

Em seguida, analisamos os sintagmas nominais com base nas estruturas FCA (Figura 7.6) geradas para aqueles cuja frequência em relações verbo-argumento era superior a 1. Considerando informações da VerbNet e observando a estrutura FCA da Figura 7.6b, o papel Agent é associado geralmente a elementos que exerçam algum tipo de controle ou influência sobre outros elementos. Esse é o caso de *technology* no trecho: *"... introduced new technology in mechanical design automation that will improve mechanical engineering productivity."* (sentença 0 do texto wsj_0055).

Já os sintagmas anotados com o papel Patient são aqueles que sofrem alguma modificação decorrente da ação dos agentes, tal como *market* em "*South Korea has opened its market to foreign cigarettes but ...*" (trecho da sentença 43 do texto wsj_0037). Como optamos por analisar apenas unigramas, alguns desses sintagmas, no entanto, ficaram pouco informativos como *use* e *condition.* Ambos fazem mais sentido como n-gramas, para $n \geq 2$, tal como *working*



(a) (b)

Figura 7.6 – Estruturas FCA para a classe VerbNet 45.4.

conditions no trecho "*... improve working conditions (better offices and more vacations, for example)...* " (trecho da sentença 11 do texto wsj_0094).

Da mesma forma que *use* e *condition*, o unigrama *demand*, quando analisado de forma isolada, é igualmente pouco significativo. Além do papel semântico Patient, ele também foi associado ao papel Theme. No entanto, no caso deste último papel, *demand* aparece apenas como argumento do verbo *to taper*. Dada a reduzida amostra do papel Theme para classe 45.4, sua análise não teve prosseguimento.

7.4 Considerações sobre este capítulo

Embora tenhamos descartado o uso da estrutura RCA em nossa pesquisa, acreditamos que tal extensão deva ser estudada em maior profundidade. Possivelmente, a inclusão de mais relações tais como do tipo substantivo-adjetivo ou mesmo substantivo-advérbio possam contribuir na qualificação dos conceitos formais. Essa suposta qualificação pode tornar a generalização de relações entre objetos para relações entre conceitos, proposta pelo método RCA, mais interessante para a construção de estruturas conceituais.

Um problema que observamos, no entanto, no uso dessa extensão é a escassez de ferramentas. Como mencionado na Seção 5.3.5, encontramos apenas uma única ferramenta que gera estruturas conceituais do tipo RCA, a ERCA.

No que se refere às classes VerbNet estudadas, observamos que elas delimitam os papéis semânticos. Conforme a classe, apenas determinados papéis são associados aos argumentos dos verbos. Por esta razão, acreditamos que as classes VerbNet podem ser um caminho de pesquisa interessante para responder às questões elencadas no aspecto "identificação das relações não taxonômicas", apresentado na Seção 6.2.

Para isso, seria necessário estudar métodos e heurísticas capazes de classificar os verbos uma vez que VerbNet não é completa, ou seja, ela não possui classes para uma grande parte dos verbos. De posse de um classificador com esse fim e com o auxílio dos etiquetadores de papéis semânticos seria possível estudarmos as relações entre papéis e sua relevância para diferentes domínios.

Quanto à questão de usar as classes de verbos para rotular as relações transversais (questão também mencionada no aspecto "identificação das relações não taxonômicas"), precisaríamos associar, às denominações numéricas atuais, rótulos textuais. Embora essa ideia seja interessante, acreditamos ser difícil sua realização de forma automática. Para que os rótulos façam sentido teriam que ser definidos conforme o domínio e, nesse caso, a sugestão ou mesmo crítica humana são mais necessárias. Por outro lado, como estudamos apenas duas classes, não conseguimos enxergar por completo a complexidade envolvendo a definição desses rótulos. É possível que a relação estabelecida entre os papéis de uma mesma classe não seja tão clara. De qualquer forma, acreditamos que caiba investigar tal possibilidade.

Continuamos nosso estudo analisando, no capítulo seguinte, a contribuição dos papéis semânticos e das classes de verbos às estruturas FCA.

8. ESTUDO II - REPRESENTAÇÃO DE INFORMAÇÕES SEMÂNTICAS EM CONCEITOS FORMAIS

Este capítulo apresenta o estudo realizado para determinar a contribuição dos papéis semânticos e das classes de verbos às estruturas do tipo FCA. Com esse fim investigamos de forma exploratória o modo como a informação sobre os papéis semânticos e classes de verbos pode ser incluída em estruturas do tipo FCA. Para isso, definimos algumas configurações de contextos formais com e sem essas informações semânticas. Essas configurações, as quais chamamos de casos de estudo, são descritas na Seção 8.1. A forma de seleção de características para construção dessas configurações, bem como o método de avaliação utilizado na análise comparativa entre essas configurações, são apresentados na Seção 8.2. Já a análise em si é comentada nas Seções 8.3, 8.4 e 8.5.

8.1 Contextos formais: casos de estudo

Nesta seção detalhamos as configurações iniciais dos contextos formais que decidimos investigar. Essas configurações foram organizadas em 6 casos de estudo. A diferença entre os casos está basicamente na escolha das informações semânticas que são usadas e, ainda, na forma como essas informações são representadas.

O objetivo é, por meio desses casos, analisar a contribuição dos papéis semânticos e das classes dos verbos para a construção dos conceitos formais e também estabelecer, entre as estudadas, a representação que provê uma estrutura conceitual com conceitos mais coesos.

Para viabilizar a comparação das estruturas FCA geradas a partir dos contextos definidos por esses casos, usamos, em todos eles, o mesmo conjunto G de objetos. Com o objetivo de facilitar o entendimento das configurações desses contextos formais, utilizamos, nos exemplos abaixo, as relações entre verbos e respectivos argumentos constantes na Tabela C.6 do Apêndice C. Assim sendo, em nossos exemplos, o conjunto G corresponde aos argumentos *{share, shareholder, stockholder, company, dividend, analyst}*. Cabe mencionar que, em G, há apenas os sintagmas nominais constituídos por substantivos comuns.

Descrevemos, a seguir, os 6 casos[1] estudados:

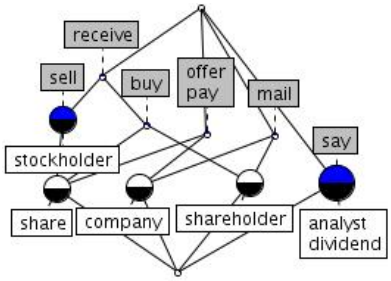
- caso $1_{(sn,v)}$: Neste caso, a estrutura FCA é construída da forma mais "tradicional", seguindo trabalhos como o de Cimiano em [41]. Logo, essa estrutura não inclui qualquer informação semântica. Para definir seu contexto formal, usamos os verbos, correspondentes aos sintagmas nominais em G, como atributos formais, construindo, assim, o conjunto M_{caso1}=*{receive, mail, ...}*. Este caso foi definido para fins de comparação. A meta era comparar a estrutura conceitual gerada a partir dele com as geradas pelos demais casos, que incluem, em suas representações, informações semânticas. Desta forma, poderíamos medir a contribuição da classe dos verbos e dos papéis semânticos na definição dos conceitos formais. A Figura 8.1 apresenta um exemplo de contexto formal e da respectiva estrutura FCA para este caso de estudo.

- caso $2_{(sn,psV)}$: Para analisarmos a inclusão de papéis semânticos em estruturas FCA, definimos o caso 2 e o caso 3. No caso 2, tomamos como exemplo o trabalho de Rudolf Wille

[1] Junto a cada caso, especificamos uma legenda em subscrito que identifica a relação de incidência I da estrutura FCA. Nessas legendas, utilizamos abreviações: *sn* corresponde a sintagma nominal; *v*, a verbo; *psV*, a papel semântico VerbNet; e *cV*, a classe VerbNet.

	atributos						
objetos	receive	mail	buy	sell	say	pay	offer
share	x		x	x		x	x
shareholder	x	x	x				
stockholder	x			x			
company		x				x	x
dividend					x		
analyst					x		

(a)



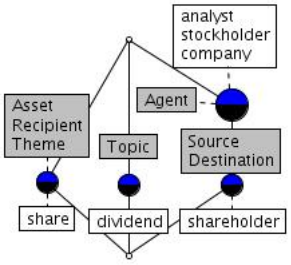
(b)

Figura 8.1 – Contexto formal e estrutura FCA para o caso $1_{(sn,v)}$.

[219] que utiliza os papéis semânticos como se fossem classes. Apesar de melhor entendermos papéis semânticos como relações transversais entre conceitos, experimentamos a proposta de Wille e definimos os papéis semânticos VerbNet, correspondentes aos sintagmas de G, como atributos. Desta forma, no contexto formal deste caso, o conjunto de atributos é formado por M_{caso2}={*Agent*,*Theme*,...,}. Um exemplo desse contexto formal e da estrutura FCA correspondente são apresentados na Figura 8.2.

	atributos						
objetos	Agent	Theme	Recipient	Destination	Source	Topic	Asset
share		x	x				x
shareholder	x			x	x		
stockholder	x						
company	x						
dividend						x	
analyst	x						

(a)



(b)

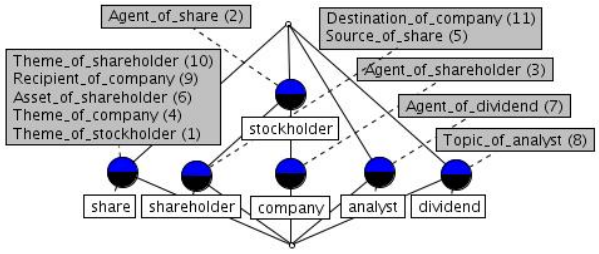
Figura 8.2 – Contexto formal e estrutura FCA para o caso $2_{(sn,psV)}$.

- caso $3_{(sn,psV\_sn)}$: Analisamos também a inclusão dos papéis semânticos como relações transversais. Investigamos, neste caso, o uso de contextos lexicossemânticos como atributos. Esses contextos foram inspirados nos contextos lexicossintáticos usados por Otero *et al.* em [157]. Os contextos foram definidos a partir de relações de incidência do tipo *(noun, VerbNetSemanticRole_of_noun)*. Para isso, consideramos apenas os verbos e argumentos que aparecem em uma mesma sentença de um texto. Por exemplo, para a tupla[2] *analyst(Agent)-say-dividend(Topic),* formada pelo verbo *to say* e seus argumentos (extraídos da sentença 5 do texto wsj_0090), são atributos de M_{caso3}: *agent_of_dividend* e *topic_of_analyst*. Logo, as relações $G \times M_{caso3}$ *(analyst, agent_of_dividend)* e *(dividend, topic_of_analyst)* são elementos da relação de incidência I_{caso3}. A Figura 8.3 apresenta um exemplo desse contexto formal e de sua respectiva estrutura FCA.

- caso $4_{(sn,cV)}$: Para analisarmos a influência das classes dos verbos construímos, então, a estrutura FCA, utilizando como atributos as classes VerbNet dos verbos, formando, assim o conjunto M_{caso4} = *{13.5.2, 11.1-1, ...}*. A Figura 8.4 apresenta um exemplo de contexto formal e da respectiva estrutura FCA para este caso de estudo.

objetos	atributos										
	1	2	3	4	5	6	7	8	9	10	11
share	x			x		x			x	x	
shareholder		x			x						x
stockholder		x									
company		x	x								
dividend								x			
analyst							x				

(a)

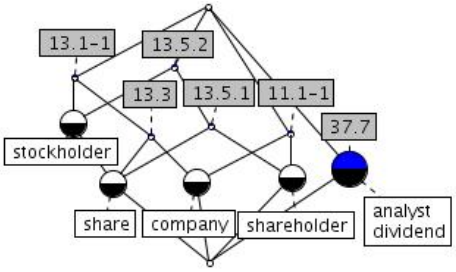


(b)

Figura 8.3 – Contexto formal e estrutura FCA para o caso $3_{(sn,psV\_sn)}$.

objetos	atributos					
	13.5.2	11.1-1	13.5.1	13.1-1	37.7	13.3
share	x		x	x		x
shareholder	x	x	x			
stockholder	x			x		
company		x		x		x
dividend					x	
analyst					x	

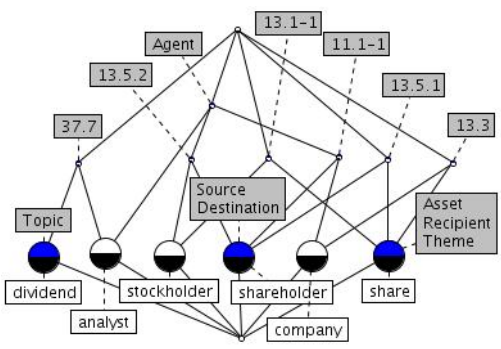
(a)



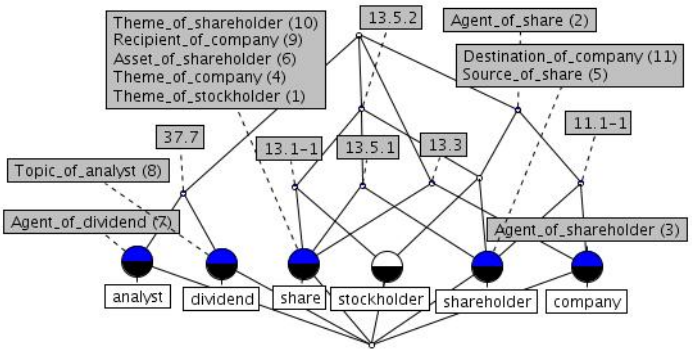
(b)

Figura 8.4 – Contexto formal e estrutura FCA para o caso $4_{(sn,cV)}$.

- caso $5_{(sn,psV)+(sn,cV)}$: Estudamos, neste caso, a contribuição dos papéis semânticos em conjunto com as classes de verbos. Como investigamos duas maneiras de representar os papéis semânticos (casos 2 e 3), usamos inicialmente a forma definida no caso 2. Assim, definimos o conjunto de atributos para este caso como M_{caso5}=$\{M_{caso2} \cup M_{caso4}\}$. A Figura 8.5a mostra um exemplo de estrutura FCA para este caso.



(a)



(b)

Figura 8.5 – Estruturas FCA para o casos $5_{(sn,psV)+(sn,cV)}$ e o caso $6_{(sn,psV\_sn)+(sn,cV)}$.

- caso $6_{(sn,psV\_sn)+(sn,cV)}$: Neste caso também analisamos contribuição dos papéis semânticos VerbNet em conjunto com as classes de verbos. No entanto, usamos a forma definida

[2]Esta tupla é um dos exemplos apresentados na Tabela C.6 do Apêndice C.

no caso 3. Logo, o conjunto de atributos para este caso corresponde a M_{caso6}=$\{M_{caso3} \cup M_{caso4}\}$. A Figura 8.5b mostra um exemplo de estrutura FCA para este caso.

Na seção seguinte, apresentamos a forma de seleção das relações incluídas em nosso estudo, bem como as medidas de avaliação estrutural utilizadas.

8.2 Seleção e avaliação

Antes de iniciarmos a avaliação dos casos de estudo mencionados, analisamos algumas formas de seleção a serem aplicadas às relações entre os verbos e seus argumentos. Nosso objetivo era escolher, dentre essas relações, as quais estruturamos na forma de tuplas[3], aquelas cujas informações fossem mais representativas para o domínio de Finanças.

Estabelecemos, então, 8 formas de seleção distintas. Em algumas formas, não usamos sementes; e, em outras, as usamos para iniciar o processo de recuperação efetuado pelo "operador mais". Variamos, também, os pontos de corte, usando valores de 2 a 5.

Nas configurações que não envolviam sementes, os pontos de corte foram usados para determinar a frequência mínima dos argumentos (sintagmas nominais) existentes nas relações. Nas que eram direcionadas por sementes, os pontos de corte restringiram o processo de recuperação do "operador mais". Ele se limitava a buscar tuplas em que os argumentos relacionados às sementes possuíam a frequência mínima indicada pelo corte. Usamos como sementes iniciais aquelas definidas na Seção 6.4.3. Cabe mencionar ainda que, durante a análise dessas configurações de seleção, não consideramos as informações semânticas existentes nas tuplas.

Uma vez definidas as listas de termos resultantes de cada processo de seleção, comparamos seus termos aos nomes das classes existentes na ontologia LSDIS Finance (descrita na Seção 5.2). Usamos para isso a medida estrutural CMM (Seção 2.6.1), que avalia a cobertura de uma ontologia para um conjunto de termos previamente informado. Ela contabiliza e pondera casamentos exatos e parciais entre os rótulos das classes e os termos informados. No cálculo dessa métrica, usamos os pesos sugeridos pelos autores Alani e Brewster em [2], que aplicam o peso $0,6$ para casamentos exatos e $0,4$ para parciais.

Cabe ressaltar que não aplicamos a medida CMM também para a ontologia Finance (descrita na Seção 5.2), pois na época em que esse estudo foi realizado ainda não tinhamos encontrado tal ontologia. Além disso, julgamos a quantidade de classes disponíveis na LSDIS Finance adequada para essa investigação.

Na Tabela 8.1 apresentamos as formas de seleção que testamos e os valores calculados para a métrica $\text{CMM}_{\text{LSDIS}}$ em relação à ontologia LSDIS Finance. Decidimos aplicar as configurações de seleção que obtiveram os 4 melhores valores para $\text{CMM}_{\text{LSDIS}}$: 1, 2, 3 e 5. A configuração 1, que utiliza como ponto de corte o valor 2 e não faz uso de sementes, foi a que obteve maior quantidade de casamentos com os termos da ontologia testada. Considerando a totalidade de casamentos de termos (exatos+parciais), a ontologia LSDIS Finance para tal configuração obteve cerca de 53% elementos coincidentes.

Definidas as formas de seleção, precisavámos de uma medida que nos permitisse avaliar a contribuição das informações semânticas no que se refere à definição dos conceitos formais de uma estrutura FCA.

Como comentamos na Seção 2.6, a avaliação de estruturas conceituais, embora bastante pesquisada, ainda não é um tema consolidado. Quando avaliamos estruturas baseadas em FCA, aumentam as dificuldades pois esse tipo de investigação é mais recente. Encontramos apenas duas medidas para esse tipo de avaliação, as quais foram estudadas na Seção 3.7.

Visto que nossa meta era analisar os conceitos formais sob um olhar semântico, das duas medidas estudadas, apenas a medida Sim (Seção 3.7.2) atende a nosso propósito. Entretanto,

[3]As tuplas seguem o formato apresentado na Tabela C.6 do Apêndice C.

Tabela 8.1 – Formas de seleção e valores da métrica estrutural *CMM* para a ontologia LSDIS Finance

N.	tipo	corte	exatos	parciais	$\mathbf{CMM_{LSDIS}}$
1	s/ sementes	2	18	125	**60,8**
2	s/ sementes	3	13	109	**51,4**
3	s/ sementes	4	11	97	**45,4**
4	s/ sementes	5	10	88	41,2
5	10 sementes	2	14	98	**47,6**
6	10 sementes	3	7	61	28,6
7	10 sementes	4	5	41	19,4
8	10 sementes	5	5	41	19,4

ela o atende de forma parcial, pois os atributos dos contextos formais de nossos casos de estudo são de tipos diferentes, difíceis de serem comparados. A dificuldade de comparar, por exemplo, a classe de um verbo com um papel semântico, inviabilizou o uso dessa medida.

Focamos nossa análise, então, nos objetos formais. Precisávamos de uma medida de ordem estrutural que avaliasse semanticamente os grupos de objetos de um conceito formal. Assim, poderíamos verificar quais configurações geravam agrupamentos cuja relação semântica entre os objetos era mais representativa. Usamos, para isso, a medida de similaridade semântica, SSM (Seção 2.6.1), que calcula quão próximos estão, em uma determinada ontologia, os conceitos que casam exatamente ou parcialmente com termos informados. Como a métrica foi aplicada aos objetos formais de cada conceito das estruturas FCA analisadas, a SSM acabou funcionando como uma espécie de medida de coesão lexical.

Halliday e Hasan [90] usam o termo coesão para se referir "às relações de significado que existem dentro de um texto". Segundo esses autores, a coesão ocorre quando a interpretação de um elemento é dependente de outro elemento do discurso. É expressa tanto por meio da gramática quanto do vocabulário. Sendo neste último caso chamada de coesão lexical, a qual analisa a relação semântica entre as palavras do texto [205]. A coesão lexical toma como base relações como sinomínia, hiponímia, meronímia e antonímia para determinar as relações de sentido entre as palavras do texto [90].

Trabalhos como o de Teike e Fankhauser [205], ainda que com fim diferente, usam a WordNet para medir a coesão lexical. Teike e Fankhauser têm como objetivo auxiliar a anotação de textos identificando automaticamente n-gramas cujos elementos estão mais fortemente relacionados. A coesão lexical é determinada com base no comprimento do menor caminho, existente na hierarquia WordNet, entre os *synsets* dos termos sob análise.

A exemplo de trabalhos desse tipo, empregamos uma medida, comumente aplicada à WordNet, para determinar a coesão lexical dos objetos de um conceito formal. A medida escolhida foi a definida por Wu e Palmer [224]. A decisão por sua utilização se baseia em duas razões. A primeira é que ela foi mencionada, pelos autores de SSM em [2], como uma das medidas que poderiam ser usadas no cálculo da métrica. Outra razão é que o pacote NLTK (Seção 5.3.2) dispõe da implementação de tal medida para WordNet e poderiamos utilizá-la facilmente. Cabe mencionar ainda que a medida gera valores normalizados ($[0; 1]$), o que facilita a sua interpretação.

Optamos também por aplicar a métrica SMM em relação à ontologia LSDIS Finance e essa decisão igualmente teve duas razões. A primeira é que, apesar da extensão e da riqueza em relações da base WordNet, tais relações não se referem a um domínio específico. Como esse é o nosso caso, imaginamos que a medida Wu e Palmer, aplicada à estrutura WordNet, poderia não capturar a relação semântica esperada e gerar valores menos expressivos. A segunda é que, embora na ontologia LSDIS Finance, o conjunto de conceitos seja menor, é mais usual

conceitos rotulados com n-gramas ($n > 1$) e as relações entre esses conceitos são de domínio. Esses fatores podem gerar resultados semanticamente mais significativos quanto à qualidade dos agrupamentos (conceitos).

No caso da ontologia LSDIS Finance, além de implementar a métrica SSM, tivemos que codificar também o cálculo da medida de Wu e Palmer para tal estrutura conceitual. Essa implementação foi realizada em Java e segue as fórmulas apresentadas a seguir. A medida SSM_E, expressa na Equação 6.1, indica a coesão lexical média dos N conceitos de uma estrutura FCA em relação a uma estrutura conceitual E. Já a medida ssm_i (Equação 6.2), calcula a similaridade do conjunto de objetos G de um conceito i de uma estrutura FCA, com base na medida de Wu e Palmer (wup). Caso esse conjunto G possua cardinalidade igual a 1, a medida ssm_i resulta zero. Por fim, a medida wup_E, apresentada na Equação 6.3, estima a similaridade entre os conceitos c_1 e c_2 em uma estrutura E. Nessa equação, a corresponde ao ancestral comum e mais específico dos conceitos c_1 e c_2; p, à profundidade de um nodo qualquer, ou seja, o comprimento do caminho (em nós) desse nodo ao nodo raiz; e d, à menor distância (em nós) de c_1 a c_2.

$$SSM_E = \frac{1}{N} \sum_{i=1}^{N} ssm_i \tag{6.1}$$

$$ssm_i = \begin{cases} \dfrac{1}{|G_i|} \displaystyle\sum_{j=1}^{|G_i|-1} \sum_{k=j+1}^{|G_i|} wup_E(o_j, o_k) & para \quad |G_i| > 1 \quad e \quad o_j, o_k \in G_i \\ 0 & em \qquad c.c. \end{cases} \tag{6.2}$$

$$wup_E(c_1, c_2) = \begin{cases} \dfrac{2 \times p(a(c_1, c_2))}{d(c_1, a(c_1, c_2)) + d(c_2, a(c_1, c_2)) + 2 \times p(a(c_1, c_2))} & para \quad c_1, c_2 \in E \\ 0 & c.c. \end{cases} \tag{6.3}$$

Analisamos também a relação entre a cardinalidade do conjunto de atributos de cada estrutura com a quantidade de conceitos formais produzidos e, ainda, a altura e a largura dessas estruturas. Outro elemento avaliado é a quantidade de arestas dessas estruturas: quanto maior for esse valor, maior será a complexidade na construção do reticulado correspondente.

Nas seções seguintes são apresentadas as análises realizadas para os casos de estudos descritos na Seção 8.1. Cabe comentar, que para a realização desse estudo, geramos várias estruturas FCA. Os grafos referentes a essas estruturas foram implementados em linguagem Java. Usamos para isso o algoritmo *Bordat* (comentado na Seção 3.5.2), o qual foi escolhido pela sua simplicidade de implementação. Como já mencionado, também implementamos em Java a leitura da ontologia citada e a maioria das medidas estruturais utilizadas na avaliação das estruturas FCA. A ferramenta Concept Expert (Seção 5.3.5) foi utilizada apenas para visualização dos reticulados de conceitos.

8.3 Análise I : estudo preliminar

Nesta seção apresentamos o primeiro estudo comparativo referente aos 6 casos apresentados na Seção 8.1. Nessa análise[4], usamos as medidas estruturais e as 4 formas de seleção escolhidas na seção anterior. Ao gerar os contextos formais dos casos, variamos ainda os papéis semânticos VerbNet considerados nesses contextos. Nosso objetivo era investigar a influência dos papéis semânticos na formação dos conceitos formais. Analisamos duas configurações com esse fim. Uma incluindo apenas os 4 papéis semânticos mais frequentes (Agent, Theme, Patient e Topic)

[4]Dados complementares a essa análise são apresentados no Apêndice D.

e outra considerando todos os papéis semânticos disponíveis nas tuplas selecionadas (entre 18 e 20 papéis).

A Tabela 8.2 descreve os resultados referentes à medida estrutural SSM para os 6 casos estudados. Esses resultados foram produzidos a partir da forma de seleção 1, na qual não usamos sementes e realizamos o ponto de corte com o valor 2. Além disso, os contextos formais desses casos foram gerados considerando apenas os 4 papéis semânticos mais frequentes. Nessa tabela são apresentados dados quanto ao número de objetos e atributos de cada contexto formal analisado. Incluímos também na tabela, a quantidade de conceitos formais gerados e as medidas SSM calculadas. A medida SSM_W corresponde à coesão lexical relativa à base WordNet; e a SSM_L, à coesão lexical relativa à ontologia LSDIS Finance. A última coluna da tabela apresenta a média aritmética dessas 2 medidas.

Tabela 8.2 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 1: sem sementes, corte 2 (4 papéis semânticos).

caso	relação I (g,m)	#objetos	#atributos	#conceitos	SSM_W	SSM_L	média
1	(sn,v)	178	71	119	0,21	0,11	0,16
2	(sn,psV)	178	4	12	0,48	0,47	**0,48**
3	(sn,psV_sn)	178	214	151	0,09	0,05	0,07
4	(sn,cV)	178	45	99	0,23	0,13	0,18
5	(sn,psV)+(sn,cV)	178	49	196	0,29	0,23	0,26
6	(sn,psV_sn)+(sn,cV)	178	259	247	0,15	0,12	0,14

Observando-se os dados da Tabela 8.2, podemos perceber que apenas os casos 3 e 6, que contêm a relação (sn, psV_sn), obtiveram, na média, coesão lexical inferior à do caso $1_{(sn,v)}$. O caso $3_{(sn,\ psV_sn)}$ foi, inclusive, o que gerou conceitos com mais baixa coesão. Isso aconteceu em decorrência da especificidade dos atributos da forma *VerbNetSemanticRole_ of_ noun* (psV_-sn). Poucos objetos compartilhavam tais atributos, o que acabou produzindo muitos conceitos cuja cardinalidade do conjunto de objetos era 1. A Tabela 8.3, que apresenta os demais dados estruturais (número de arestas, altura e largura média da estrutura FCA) para mesma forma de seleção e quantidade de papéis, mostra que 58,3% dos conceitos formais gerados para o caso 3 possuíam conjuntos unitários de objetos. Na tabela chamamos a esses conceitos simplesmente de "unitários".

Tabela 8.3 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 1: sem sementes, corte 2 (4 papéis semânticos)

caso	relação I (g,m)	#arestas	#unitários (%)	altura	largura
1	(sn,v)	237	35 (29,4)	5	56
2	(sn,psV)	20	0	4	4
3	(sn,psV_sn)	271	88 (58,3)	5	100
4	(sn,cV)	211	34 (34,3)	5	40
5	(sn,psV)+(sn,cV)	471	52 (26,5)	7	61
6	(sn,psV_sn)+(sn,cV)	519	106 (42,9)	6	122

O caso 6, além de apresentar uma coesão mais baixa que o 1, também apresentou outro inconveniente que foi a quantidade de arestas (Tabela 8.3). Ele possuía aproximadamente o dobro das arestas do caso 1. A quantidade bem maior de atributos do caso 6 em relação ao 1, aumentou a complexidade do reticulado e, consequentemente, o processamento computacional.

O caso 5 apresentou o mesmo problema. Apesar de conter cerca de 30% de atributos a menos que o 1, também produziu muitas arestas. Por outro lado, percebe-se que a combinação de

classes de verbos e papéis semânticos resultou em uma estrutura conceitual mais especializada. Dentre as estruturas, a do caso 5 é a de maior altura: 7.

Analisando-se os casos 4, 5 e 6, podemos notar que a presença das classes de verbos (cV) como atributos melhora a coesão lexical. Comparando-se o caso 4 com o 1, percebe-se que, além da melhora na coesão, houve uma redução: no número total de conceitos (maior agrupamento), no número de conceitos com conjunto unitário de objetos (melhor agrupamento) e na quantidade de arestas (menor processamento).

De todos os casos analisados, o caso 2 foi o que obteve maior índice de coesão lexical. No entanto, concentrou os objetos em poucos conceitos. A generalidade de seus atributos, que são papéis semânticos (psV), deve ser a razão da alta coesão. A presença de mais objetos no mesmo conceito formal aumenta a quantidade de combinações de pares de objetos que são submetidos à medida de similaridade de Wu e Palmer. Visto que esses objetos possuem alguma relação semântica (no mínimo a definida pelo próprio papel semântico), a similaridade resultante acaba sendo maior.

No entanto, ao incluirmos todos os papéis semânticos nos contextos formais dos casos de estudo, mesmo mantendo a forma de seleção, percebemos uma redução na coesão lexical do caso 2 (Tabela 8.4). Essa queda na coesão deve estar relacionada ao acréscimo de atributos (de 4 para até 20 papéis semânticos) e à baixa frequência da maioria deles. Em geral, mais atributos provocam uma maior distribuição dos objetos em conceitos. Nesse caso especificamente, aumentou não apenas a quantidade de conceitos mas também a de conceitos com conjunto unitário de objetos, que foi para 25% (Tabela D.7 do Apêndice D).

Tabela 8.4 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 1: sem sementes, corte 2 (todos os papéis)

caso	**relação I (g,m)**	**#objetos**	**#atributos**	**#conceitos**	**SSM_W**	**SSM_L**	**média**
1	(sn,v)	377	153	289	0,21	0,11	0,16
2	(sn,psV)	377	20	84	0,32	0,17	0,25
3	(sn,psV_sn)	377	579	356	0,09	0,04	0,07
4	(sn,cV)	377	86	261	0,23	0,13	0,18
5	(sn,psV)+(sn,cV)	377	106	588	0,32	0,20	**0,26**
6	(sn,psV_sn)+(sn,cV)	377	665	604	0,20	0,12	0,16

Já para o caso 5 tal inclusão resultou em uma melhor coesão lexical, ainda que com os mesmos problemas relatados anteriormente. Tais variações na coesão lexical não são decorrentes apenas da inclusão de mais papéis semânticos mas também do tipo de seleção utilizada.

Para realizarmos, então, uma análise mais abrangente construímos alguns gráficos para avaliar determinados aspectos das estruturas conceituais geradas a partir dos 6 casos. Iniciamos estudando a medida SSM. Para isso, criamos dois gráficos usando as médias SSM apresentadas nas tabelas desta seção e nas tabelas do Apêndice D. O gráfico da Figura 8.6 mostra o comportamento da média SSM para os 6 casos. Nesse gráfico, os contextos formais foram gerados para as 4 formas de seleção e incluem apenas os 4 papéis semânticos mais frequentes. Já no gráfico da Figura 8.7, os contextos formais incluem todos os papéis semânticos encontrados nas tuplas selecionadas.

Analisando-se esses gráficos, observamos que o caso $3_{(\text{sn,ps}V_s\text{n})}$ independentemente da forma de seleção e número de papéis, gerou a coesão lexical mais baixa. Da mesma forma, o caso $6_{(\text{sn,ps}V_s\text{n})+(\text{sn,cV})}$ também não apresenta médias que superem as do caso $1_{(\text{sn,v})}$, embora inclua mais informações semânticas em seus contextos formais. Essa constância se repete igualmente para o caso $2_{(\text{sn,pV})}$ que, mesmo com tais variações, mantém seus índices de coesão lexical superiores aos do caso $1_{(\text{sn,v})}$.

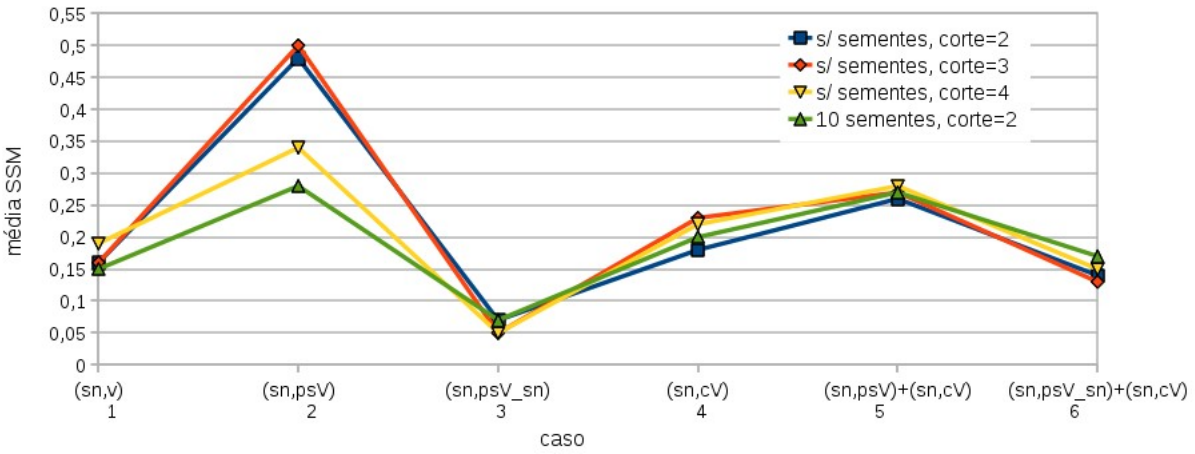


Figura 8.6 – Comparação das medidas SSM considerando 4 papéis semânticos.

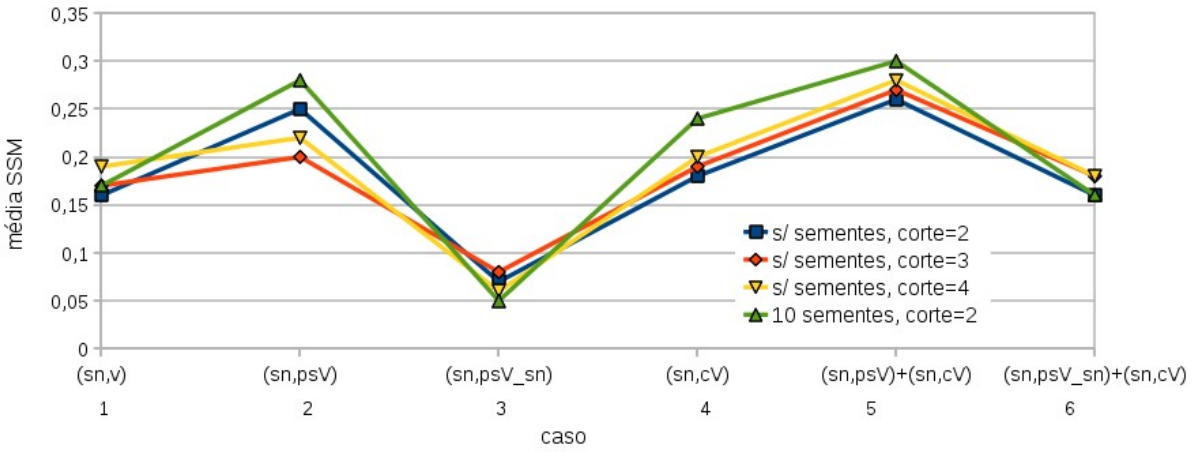


Figura 8.7 – Comparação das medidas SSM considerando todos os papéis semânticos.

O caso $4_{(sn,cV)}$ e o caso $5_{(sn,psV)+(sn,cV)}$, apesar de gerarem conceitos mais coesos que o caso $1_{(sn,v)}$, ao terem seus contextos alterados pela inclusão de mais papéis semânticos, para maioria das formas de seleção, não apresentam melhora significativa da medida SSM. No entanto, para forma de seleção 5 (10 sementes, corte 2), esses casos produzem resultados acima dos demais na presença de uma variedade maior de papéis em seus contextos formais. Visto que ambos os casos incluem classes de verbo e esta forma de seleção utiliza o "operador mais", é possível que, ao incluirmos mais papéis, esse operador tenha recuperado um número maior de tuplas cujos verbos possuem mais classes em comum.

Ainda com o fim de analisar os agrupamentos, observamos também os percentuais de conceitos com conjuntos unitários de objetos. Para isso, geramos os gráficos das Figuras 8.8 e 8.9 que correspondem, respectivamente, a configurações com os 4 papéis mais frequentes e com todos os encontrados nas tuplas selecionadas. Percebemos que o comportamento dos casos em ambos os gráficos é muito semelhante. A redução do número de papéis afeta significativamente apenas o caso $2_{(sn,psV)}$. Esse caso na presença de um número menor de papéis, que são seus atributos, gera um percentual menor ou nulo de conceitos com conjuntos unitários de objetos.

Entretanto, de maneira geral, o que influencia diretamente a quantidade de unitários é a forma de seleção utilizada. Em ambas as situações, a forma de seleção 5, que utiliza sementes, produziu um menor índice de unitários.

Com base nas tabelas do Apêndice D, complementamos a observação de que independentemente do número de papéis e da forma de seleção utilizados, os casos 5 e 6 continuam, indesejavelmente, a gerar estruturas FCA com uma maior quantidade de arestas.

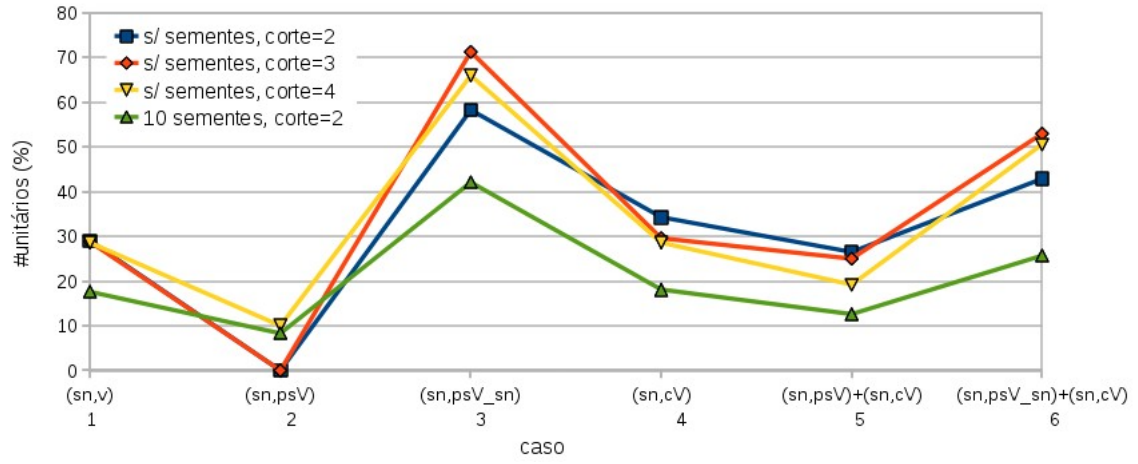


Figura 8.8 – Percentuais de unitários em contextos contendo 4 papéis semânticos.

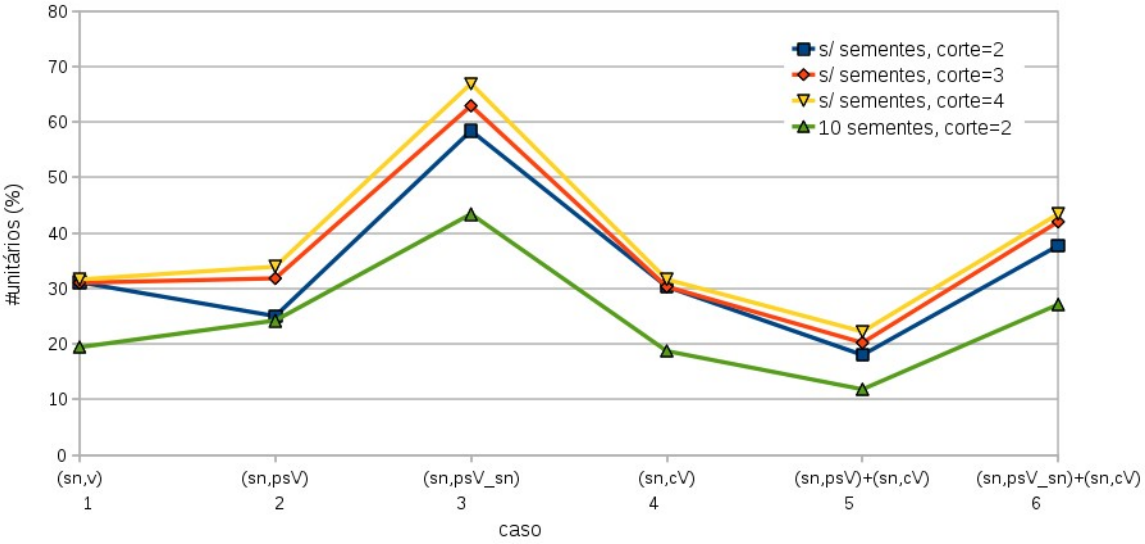


Figura 8.9 – Percentuais de unitários em contextos com todos papéis semânticos.

Cabe mencionar que analisamos também as relações não taxonômicas existentes na ontologia LSDIS Finance. Confirmando a literatura, relações desse tipo constavam em menor número nessa ontologia. Na LSDIS Finance, encontramos 50 relações funcionais e 68 relações descrevendo propriedades. Verificamos, então, se os conceitos formais provenientes dos casos estudados coincidiam com tais relações. Não encontramos nenhum casamento "perfeito" entre os conceitos formais e essas relações.

8.4 Análise II: estudo de heurísticas

Com o objetivo de tentar melhorar os resultados, principalmente do caso $3_{(\text{sn, psV\_sn})}$, incluímos algumas heurísticas no pré-processamento dos contextos formais dos casos estudados.

A exemplo do trabalho de Otero *et al.* em [72], que inspirou a configuração do caso $3_{(\text{sn, psV\_sn})}$, aplicamos uma heurística para agrupar atributos similares, baseada no coeficiente Dice. Essa medida é apresentada na Equação 6.4, onde f corresponde à frequência absoluta, min determina o menor de 2 valores e n é a quantidade de objetos compartilhados pelos atributos 1 e 2.

$$Dice(atributo_1, atributo_2) = \frac{2 \times \sum_{i=1}^{n} min(f(objeto_i, atributo_1), f(objeto_i, atributo_2))}{f(atributo_1) + f(atributo_2)} \quad (6.4)$$

Para cada atributo analisado, são geradas novas relações a partir dos seus k vizinhos (atributos mais semelhantes de acordo com a medida Dice). Otero *et al.* utilizaram k=*5*. Nós testamos os valores 4, 5 e 6 para k. Estabelecemos como similares aqueles atributos cuja medida gerou valores do intervalo $(0,5; 1)$. As novas relações foram definidas a partir dos objetos que não eram compartilhados originalmente pelos atributos considerados semelhantes. No caso de os atributos 1 e 2 serem similares, ao existir a relação $(objeto_j, atributo_1)$ em que o $objeto_j$ não é compartilhado pelo $atributo_2$, a relação $(objeto_j, atributo_2)$ é criada.

Aplicamos essas heurísticas apenas aos 4 primeiros casos de estudo. Desconsideramos os casos 5 e 6 por terem produzido muitas arestas em suas estruturas FCA. Grandes quantidades de arestas são indesejáveis computacionalmente. Em todos os testes realizados usamos a forma de seleção 5, que usa utiliza sementes e vários papéis semânticos, pois, de acordo com os resultados apresentados na seção anterior, foi a que gerou menor quantidade de unitários. Excluímos ainda atributos menos frequentes. Testamos 3, 4 e 5 como valores de corte para m, o qual corresponde à frequência mínima exigida para os atributos.

Apesar de as heurísticas terem melhorado a medida SSM de todos os 4 casos, a heurística de agrupamento baseada no coeficiente Dice foi mais efetiva para o caso $3_{(sn,\ psV\_sn)}$(Tabela E.3 do Apêndice E). Para os demais casos, os valores de k não foram tão decisivos para os resultados de SSM. Para esses casos o que prevaleceu foram os pontos de corte. A Tabela 8.5 mostra os dados das melhores médias SSM obtidas para os 4 casos após a aplicação das heurísticas. Todos os dados apresentados foram extraídos das tabelas do Apêndice E e correspondem à configuração em que $k = 4$ e $m = 5$. Essa configuração foi a que gerou resultados mais significativos para todos os 4 casos.

Tabela 8.5 – Melhores médias SSM para os casos 1, 2, 3 e 4 após aplicação de heurísticas.

caso	**relação I (g,m)**	**#objetos**	**#conceitos**	**%unitários**	**SSM_W**	**SSM_L**	**média**
1	(sn,v)	215	109	17,4	0,39	0,18	0,29
2	(sn,psV)	263	53	20,8	0,33	0,25	0,29
3	(sn,psV_sn)	140	28	21,4	0,15	**0,28**	0,22
4	(sn,cV)	234	114	19,3	0,39	0,26	**0,33**

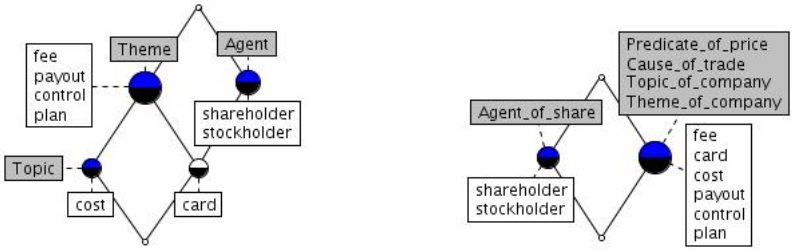
Após a aplicação das heurísticas, todos os 4 casos melhoraram as médias SSM. A melhora mais expressiva foi a do caso $3_{(sn,\ psV\_sn)}$. É interessante observar que, mesmo ainda com a menor média SSM, o caso $3_{(sn,\ psV\_sn)}$ gerou conceitos tão densos quanto o caso $2_{(sn,\ psV)}$ em número de objetos (cerca de 5 por conceito). Outro aspecto a se considerar é que a coesão lexical dos conceitos desse caso, principalmente em relação às ontologias de domínio, melhorou significativamente. Dentre todos os casos, ele obteve o valor SSM mais alto para a ontologia LSDIS Finance.

O caso $4_{(sn,\ cV)}$ foi o que gerou conceitos mais coesos. Observamos que tanto em relação à WordNet quanto em relação à ontologia LSDIS Finance, a medida SSM, para esse caso, obteve resultados expressivos.

O caso $1_{(sn,\ v)}$, na coesão média, foi tão bom quanto o caso $2_{(sn,\ psV\_sn)}$. As heurísticas aplicadas aumentaram de forma significativa a medida SSM em relação à WordNet. No entanto, para a ontologia de Finança, tal medida foi a mais baixa dentre os casos. Isso pode ser um indicativo de que as informações semânticas incluídas nos conceitos formais têm uma contribuição relevante na captura de relações de domínio.

Para que pudéssemos analisar qualitativamente os conceitos que incluem informações referentes a papéis semânticos, geramos pequenos exemplos para os casos 2 e 3. Os exemplos usam

a mesma configuração descrita na Tabela 8.5, no entanto restringimos seus contextos formais a objetos cujos sintagmas são unigramas e são compartilhados por ambos os casos. A Figura 8.10 exibe as estruturas FCA para os casos 2 e 3 geradas a partir desses contextos.



(a) (b)

Figura 8.10 – Estruturas FCA para o caso $2_{(\text{sn, psV\_sn})}$ e o caso $3_{(\text{sn, psV})}$ após o uso de heurísticas.

Analisando as estruturas, percebemos uma certa similaridade entre os conceitos. No entanto, os atributos do caso $3_{(\text{sn, psV\_sn})}$ (Figura 8.10b) são mais informativos por serem baseados em relações. Aparentemente, tais atributos conseguem delinear melhor a semântica do domínio por expressarem o contexto em que os papéis semânticos são aplicados. Já os atributos do caso $2_{(\text{sn, psV})}$ (Figura 8.10a) formam grupos de objetos aparentemente mais abrangentes. Talvez isso seja uma das explicações para o fato de a medida SSM, quando aplicada à WordNet, gerar valores mais altos para este caso.

Por outro lado, os unigramas *share*, *price*, *trade* e *company* que aparecem nos atributos do caso $3_{(\text{sn, psV\_sn})}$, não fazem parte do seu conjunto de objetos formais. Eles devem ter sido descartados em razão de estarem associados a atributos (contextos lexicossemânticos) menos frequentes. Na prática, ainda mais após o uso das heurísticas, os atributos do caso $3_{(\text{sn, psV\_sn})}$ não representam de fato relações entre objetos formais. Mesmo assim, ainda podemos utilizá-los para qualificar os objetos.

Percebemos também que os papéis semânticos Cause e Predicate não aparecem na estrutura FCA do caso $2_{(\text{sn, psV})}$, somente na do caso $3_{(\text{sn, psV\_sn})}$. Portanto, os casos 2 e 3 não se baseiam necessariamente nos mesmos papéis semânticos.

Na próxima seção, analisamos alguns papéis semânticos por meio de exemplos.

8.5 Análise III: estudo de papéis semânticos

Nesta seção estudamos de forma qualitativa o comportamente de alguns papéis semânticos. Nosso objetivo é avaliar a relevância dos mesmos na construção de estruturas conceituais baseadas em FCA dentro do escopo de nossa proposta.

Acreditamos que as características inerentes aos papéis semânticos possam ajudar a distinguir, classificar e, principalmente, relacionar os elementos extraídos de textos. A fim de analisar tais características, construímos estruturas FCA baseadas no caso $2_{(\text{sn, psV})}$ e no caso $3_{(\text{sn, psV\_sn})}$ os quais utilizam papéis semânticos em seus contextos formais. Geramos nossos exemplos usando como sementes dois termos frequentes nos *corpora*: *share* e *company*. A Figura 8.11 contém o primeiro exemplo que é a estrutura FCA gerada para o caso $2_{(\text{sn, psV})}$.

Analisando a Figura 8.11, percebemos que os papéis semânticos qualificam os objetos que formam os conceitos. Apesar disso, como eles são utilizados, na estrutura conceitual, de forma isolada, ou seja, dissociados de relações, tornam-se menos informativos. Por exemplo, o substantivo *investment* foi anotado tanto com o papel semântico Product quanto com o Agent. A representação provida pelo caso $2_{(\text{sn, psV})}$ não permite distinguir as situações em que *investment* comporta-se como Product daquelas em que funciona como Agent. Além disso, entre as se-

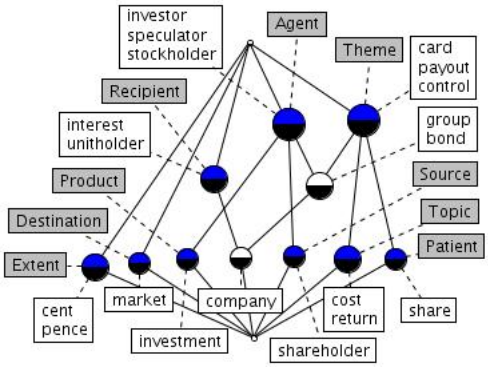


Figura 8.11 – Estrutura FCA gerada a partir das sementes *company* e *share* para o caso $2_{(\text{sn, psV})}$.

mentes *share* e *company* há relações semânticas, no entanto a estrutura apenas estabelece que o vínculo entre elas está no papel que ambas desempenham conjuntamente: Theme.

As características definidas pelos papéis semânticos aos argumentos as quais estão associados são mais evidentes quando analisamos o contexto em que eles estão inseridos. Para estudar os papéis semânticos em relações, construímos a estrutura FCA com base no caso $3_{(\text{sn, psV\_sn})}$ (nosso segundo exemplo). A Figura 8.12 apresenta tal estrutura. Cabe ressaltar que, para facilitar nossa análise, restringimos o conjunto de atributos, usando apenas aqueles que estabelecem relações com as sementes. Os atributos, portanto, seguem o formato *SemanticRole_ of_ company* e *SemanticRole_ of _share.*

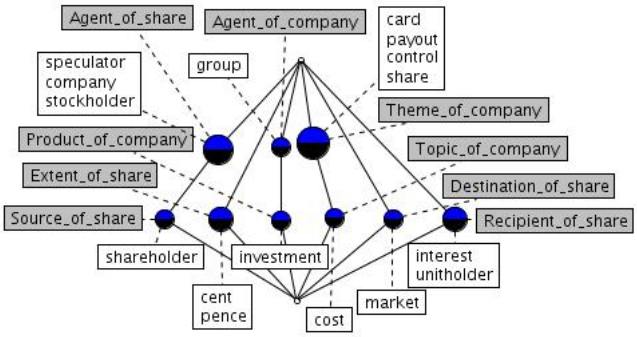


Figura 8.12 – Estrutura FCA gerada a partir das sementes *company* e *share* para o caso $3_{(\text{sn, psV\_sn})}$.

Nessa estrutura, o substantivo *investment* obviamente também está associado aos papéis Product e Agent, no entanto tais papéis possuem um contexto que é *company. Investment* é um elemento que aciona eventos (Agent) e que está relacionado a transformações ou é o resultado delas (Product) no contexto de *company.* A relação entre as sementes nessa representação é mais explícita. *Company* é um agente de *share.* E *share* é um elemento do contexto de *company* (Theme).

Há outras relações igualmente interessantes, como *unitholder* que aparece como o recipiente de *share* (Recipient). Há ainda os substantivos *speculator*, *stockholder* e *shareholder* que foram anotados como agentes, logo são interpretados como elementos que atuam sobre *share. Shareholder* também desempenha o papel de Source de *share.* Enquanto *shareholder* é a origem, *market* aparece como o destino de *share.* Os termos *cent* e *pence* também aparecem no contexto de *share*, indicando possivelmente seus valores de grandeza.

Cabe mencionar que, ao usarmos unigramas nos exemplos apresentados, alguns termos tornaram-se menos significativos, como *interest* e *group.*

Embora nosso estudo tenha sido realizado de forma pontual, visto que até agora nossa investigação tenha se restringido a um domínio (Finanças) e a um *corpus* desse domínio, acreditamos que os papéis semânticos possam, de fato, enriquecer as relações em estruturas conceituais. Claro que nem todos os papéis provêem relacionamentos significativos. Theme e Topic, por exemplo, são muito genéricos.

A incidência do papel semântico Topic, no entanto, nos parece estar mais relacionada à natureza do *corpus*, que é de cunho jornalístico, do que ao domínio em si. Nesse tipo de *corpus*, quando argumentos de um verbo em uma sentença são anotados com os papéis semânticos Agent e Topic, as chances de o argumento anotado com Agent ser uma pessoa ou uma instituição aumentam. Logo, o papel Agent pode ser usado para identificar as classes de entidades nomeadas.

A frequente presença do papel Agent nas relações identificadas entre argumentos anotados também nos parece estar mais ligada à natureza do *corpus*. Conforme o gráfico da Figura 8.13, que apresenta as 12 relações mais frequentes, Agent aparece em cerca de 50% delas.

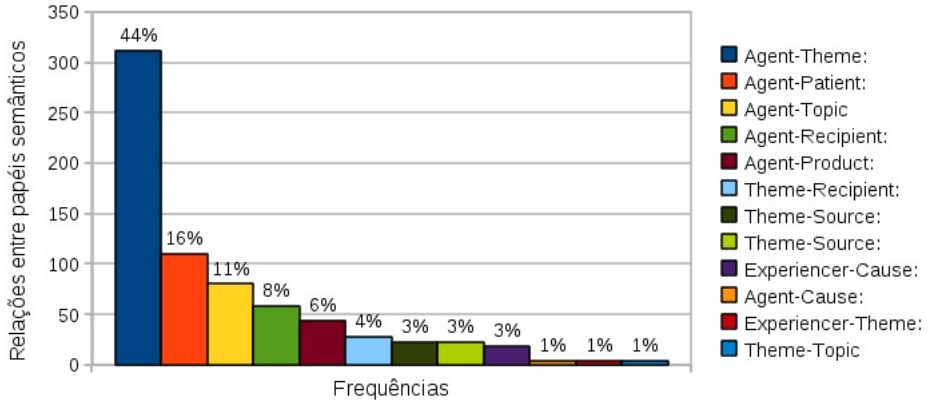


Figura 8.13 – Incidência de relações entre papéis semânticos.

8.6 Considerações sobre este capítulo

Sob o ponto de vista estrutural e lexical, observamos que a inclusão de informações semânticas nos atributos dos contextos formais, de maneira geral, teve como resultado conceitos formais mais coesos. Nesse sentido, as classes de verbos, para configuração de contexto formal proposta no caso $4_{(\mathrm{sn,cV})}$, mostraram-se mais efetivas do que verbos. As classes, além de aumentar a coesão lexical, ajudaram a reduzir a complexidade de construção do reticulado FCA, na medida em que geraram menos conceitos e arestas. Já os papéis semânticos mostraram-se mais efetivos, ainda no aspecto coesão, principalmente quando a configuração de contexto formal proposta no caso $2_{(\mathrm{sn,psV})}$ foi utilizada.

Apesar desses resultados, a interpretação das estruturas assim geradas não é tão objetiva quanto aquelas estruturas em que verbos são usados como atributos. Sob o aspecto intensional, o uso de rótulos numéricos para as classes de verbos, bem como o uso de papéis semânticos como classes e não como relações, tornam tais elementos, enquanto atributos, menos informativos do que os verbos.

Já a configuração de contexto formal descrita pelo caso $3_{(\mathrm{sn,psV\_sn})}$, na qual os papéis semânticos são utilizados como relações, apresenta atributos que nos pareceram mais descritivos intensionalmente, ainda que inicialmente (antes da aplicação das heurísticas descrita na Seção 8.4) tal configuração tenha produzido conceitos menos coesos.

Quanto às formas de seleção estudadas na Seção 8.3, o uso de sementes em conjunto com o "operador mais" mostrou-se um caminho interessante para gerar estruturas FCA. Essa aborda-

gem, além de melhorar a coesão dos conceitos, também reduziu a complexidade de construção do reticulado FCA.

Cabe mencionar que não utilizamos, em nosso estudo, contextos formais constituídos por instâncias (substantivos próprios), pois o *parser* que construímos para extrair os sintagmas nominais (apresentado na Seção 6.4.2) as gerou em um número muito reduzido.

Cabe destacar, também, que os estudos realizados nesse capítulo foram descritos em um artigo científico ainda não publicado mas com a notificação de aceite para uma conferência internacional: Language Resources and Evaluation (LREC[5] 2012).

Com o objetivo de analisar a aplicabilidade de nossa proposta, ainda para os casos de estudo estabelecidos, estudamos e apresentamos, no capítulo seguinte, a contribuição das informações semânticas na construção de conceitos formais por meio de medidas de avaliação de ordem funcional.

[5]http://www.lrec-conf.org/lrec2012/

9. ESTUDO III - APLICABILIDADE DA PROPOSTA E ESTUDOS EM LÍNGUA PORTUGUESA

Este capítulo descreve a análise realizada para determinar a aplicabilidade de nossa abordagem em relação a outros *corpora* em Língua Inglesa. Utilizamos para esse fim os *corpora* no domínio Finanças (WikiFinance) e no domínio Turismo (WikiTourism). Ambos foram extraídos do Wikicorpus 1.0 que contém textos da Wikipédia. Por meio desses *corpora* pudemos analisar nossa proposta em relação a textos de natureza mais conceitual e em domínios diferentes. Usamos esses textos para estudar a contribuição de informações semânticas na construção de conceitos formais a partir da tarefa de categorização de textos (Seção 9.1). Para avaliar os resultados nesta tarefa, usamos medidas funcionais usuais, como precisão e *recall*. Tais medidas nos permitiram comparar os resultados gerados a partir de estruturas FCA aos produzidos por ontologias de domínio e pelo algoritmo k-Nearest Neighbor (k-NN).

Ao final desse capítulo (Seção 9.2), apresentamos os estudos que realizamos para a Língua Portuguesa quanto à extração de conceitos, à categorização de textos e à construção de estruturas conceituais do tipo FCA baseadas em papéis semânticos.

9.1 Categorização de textos

De acordo com Sebastiani em [190], "categorização de textos é a tarefa de atribuir um valor lógico a cada par $(d_j, c_i) \in D \times C$, onde D é um domínio de documentos e $C = \{c_1, ..., c_{|C|}\}$ é um conjunto pré-definido de categorias". Essa tarefa pode ser definida formalmente por meio da função $\phi : D \times C \rightarrow \{V, F\}$, que descreve o comportamento de um classificador [190].

A tarefa de categorização foi utilizada para avaliar a viabilidade de uso da nossa abordagem em outros *corpora* e domínios. Convivemos com o fato de os etiquetadores de papéis semânticos serem menos precisos para domínios que não sejam Finanças e de anotarem os argumentos dos verbos com as etiquetas numéricas PropBank. A anotação dos papéis semânticos através de rótulos numéricos nos obrigou a adaptar os contextos formais dos casos de estudo propostos no Capítulo 8. Essa adaptação é descrita na Seção 9.1.1.

Cabe mencionar que escolhemos a tarefa de categorização de textos considerando os trabalhos pesquisados que utilizam o método FCA nessa área [23, 140, 177]. Baseamos nosso estudo especialmente no trabalho de Meddouri e Meddouri [140] que define regras a partir de conceitos extraídos de estruturas FCA com o objetivo de classificar instâncias em classes. Optamos por utilizar a abordagem de Meddouri e Meddouri principalmente porque poderíamos aplicá-la facilmente tanto para estruturas FCA quanto para ontologia de domínio no tange à extração das regras. E, também, porque aqueles autores mencionam que tal abordagem é mais rápida que um classificador k-NN. Por último, embora nossa meta seja avaliar as estruturas FCA de forma mais qualitativa, no que se refere à tarefa de classificação, desempenho computacional também nos interessa. Por isso, comparamos os resultados obtidos a partir de nossa proposta aos de um classificador k-NN.

Visto que o propósito de Meddouri e Meddouri é diferente do nosso, tivemos que modificar a abordagem desses autores quanto à extração de regras para que estas pudessem ser utilizadas na categorização de documentos. Essas modificações são abordadas na Seção 9.1.3.

Os *corpora* utilizados em nossa investigação e as suas divisões em conjunto de treino e de teste são o assunto da Seção 9.1.2. Já os resultados da categorização de textos realizada a partir de estruturas FCA, ontologias de domínio e de um classificador k-NN são apresentados, respectivamente, nas Seções 9.1.3, 9.1.4 e 9.1.5. Já a comparação dos resultados obtidos nessas diferentes abordagens é apresentada na Seção 9.1.6.

É importante ressaltar, também, que fizemos uso de classificadores *single-label*, os quais atribuem cada documento d_j a apenas uma categoria c_i [189]. Utilizamos essa forma de categorização visto que os textos dos *corpora* sobre os quais os classificadores atuam foram associados a apenas uma única categoria.

9.1.1 Adaptação dos contextos formais : novos casos de estudo

Para que pudéssemos utilizar as informações de *corpora* anotados com os rótulos numéricos ao estilo PropBank, tivemos que adaptar os contextos formais dos casos de estudo propostos. Como observamos que os primeiros 4 casos, analisados no capítulo anterior, produziram reticulados de conceitos de menor complexidade computacional e com coesão lexical relevante, os escolhemos para este estudo. Desta forma, procuramos adaptar apenas os contextos formais para: caso $1_{(\mathrm{sn,v})}$, caso $2_{(\mathrm{sn,psV})}$, caso $3_{(\mathrm{sn,psV\_sn})}$ e caso $4_{(\mathrm{sn,cV})}$.

O caso $1_{(\mathrm{sn,v})}$ não requereu qualquer adaptação, visto que não faz uso de informações semânticas. Ele foi mantido, como já mencionado, para fins de comparação. Já o caso $2_{(\mathrm{sn,psV})}$ não pode ser adaptado. Não podemos simplesmente substituir etiquetas de papéis semânticos VerbNet por etiquetas PropBank, pois um mesmo identificador PropBank, quando relacionado a diferentes verbos, pode ter semântica distinta. Só há garantia de uniformidade no significado de etiquetas PropBank idênticas, se os verbos que as usam pertencerem à mesma classe.

Com o objetivo, então, de viabilizar o uso dos papéis PropBank, usamos o pacote NTLK (Seção 5.3.2) para acessar a VerbNet e buscar as possíveis classes dos verbos existentes nos *corpora* utilizados nesse estudo. No entanto, tal procedimento não foi efetivo. Conseguimos associar classes somente a aproximadamente 7% das instâncias dos verbos pertencentes a esses *corpora*. Esse insucesso, decorrente possivelmente da incompletude da própria VerbNet, inviabilizou também a adaptação do caso $4_{(\mathrm{sn,cV})}$, no qual os atributos são as classes VerbNet.

Diante da ausência de classes, para a adaptação do caso $3_{(\mathrm{sn,psV\_sn})}$, nossa alternativa acabou sendo a substituição dos papéis semânticos VerbNet por papéis PropBank[1]. Diferentemente do caso $2_{(\mathrm{sn,psV})}$, os atributos do caso $3_{(\mathrm{sn,psV\_sn})}$ são definidos por contextos lexicossemânticos. Imaginamos que os sintagmas nominais usados em sua composição poderiam ajudar a distinguir o significado dos papéis ainda que seus rótulos fossem idênticos. Chamamos o contexto formal, resultante dessa adaptação, de caso $7_{(\mathrm{sn,psP\_sn})}$. A Figura 9.1 apresenta um exemplo desse contexto formal e de sua respectiva estrutura FCA. Os dados usados nesse exemplo foram retirados da Tabela C.6 do Apêndice C.

	1	2	3	4	5	6	7	8	9	10
share	x			x		x			x	x
shareholder		x		x	x					
company			x		x					
stockholder					x					
dividend							x			
analyst								x		

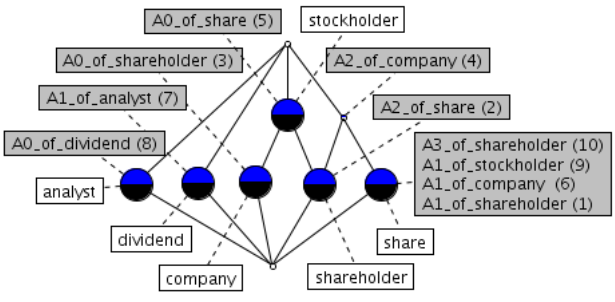


Figura 9.1 – Contexto formal e estrutura FCA para o caso $7_{(\mathrm{sn,psP\_sn})}$.

Em substituição às classes VerbNet, ainda tentamos associar aos papéis PropBank os seus respectivos verbos, constituindo contextos formais cujos atributos seguiam a forma *PropBank SemanticRole_ Verb*. Como os testes preliminares com tais contextos formais não foram satisfatórios, desistimos de usá-los.

[1] Usamos a abreviação *psP* para indicar a presença de papéis semânticos PropBank nas etiquetas em subscrito dos casos de estudo. São exemplos de etiquetas de papéis semânticos PropBank: {A0, A1, A2, ...}.

Desta forma, nos estudos sobre categorização de textos com estruturas FCA abordamos apenas: caso $1_{(sn,v)}$ e caso $7_{(sn,psP\_sn)}$.

9.1.2 Preparação dos *corpora* para a tarefa de categorização

No estudo de categorização apresentado nesse capítulo, utilizamos essencialmente textos Wikipédia extraídos do *corpus* Wikicorpus 1.0 (descrito na Seção 5.1.2). Como já comentado, utilizamos o *corpus* WikiFinance que possui 482 textos do domínio Finanças (Seção 5.1.2.1) e o *corpus* WikiTourism que contém 442 textos do domínio Turismo (Seção 5.1.2.2). Ambos foram anotados com papéis semânticos ao estilo PropBank pelo processador F-EXT-WS (Seção 5.3.4). Os termos anotados pelo processador, por meio das etiquetas POS, como adjetivos, substantivos e verbos foram, ainda, normalizados pelo lematizador TreeTagger (Seção 5.3.1).

Para viabilizar o uso dos diferentes classificadores, usamos a abordagem *train-and-test*, conforme Sebastiani em [190], e separamos os textos em conjuntos de treino e de teste. Como nosso objetivo principal era analisar a aplicabilidade de nossa proposta, não tivemos a preocupação de encontrar o melhor par de conjuntos treino e teste. Nos preocupamos apenas em garantir que os mesmos conjuntos fossem utilizados por todos os classificadores. Para que, desta forma, pudéssemos avaliar nossa abordagem e não exclusivamente a categorização em si. Escolhemos, então, arbitrariamente, em torno de 65% dos textos para formar o conjunto de treino e 35%, para o conjunto de teste. Essa separação dos textos foi realizada de forma sequencial e é apresentada na Tabela 9.1. Como pode-se observar nessa tabela, adicionalmente, ainda criamos um segundo conjunto de teste ($teste_{Wiki+PTBS}$), unindo as amostras de treino do WikiTourism aos textos Penn TreeBank Sample (PTB Sample).

Criamos esse segundo conjunto de teste para analisar o comportamento dos classificadores, especialmente o que utiliza a nossa proposta, em *corpora* de natureza distinta.

Tabela 9.1 – Separação dos textos em conjunto de treino e teste.

	WikiFinance		WikiTourism		PTB Sample	
Conjunto	~65%	~35%	~65%	~35%	100%	#
$treino_{Wiki}$	322	-	284	-	-	606
$teste_{Wiki}$	-	160	-	158	-	318
$teste_{Wiki+PTBS}$	-	-	-	158	199	357

Na seção seguinte, detalhamos o processo de categorização de textos baseado em estruturas FCA.

9.1.3 Categorização de textos baseada em regras compostas por conceitos formais

Para realizar o processo de categorização dos textos no escopo de nossa abordagem, iniciamos gerando estruturas FCA para os domínios escolhidos. Para construir tais estruturas, extraímos dos textos dos conjuntos de treino de cada domínio, as tuplas (no formato exemplicado na Tabela C.6 do Apêndice C) necessárias para definir os respectivos contextos formais. Os sintagmas nominais que integram essas tuplas foram definidos utilizando as heurísticas comentadas na Seção 6.4.2. Ao constituir as tuplas, consideramos apenas relações verbo-argumento com frequência[2] mínima igual a 2. Determinadas as tuplas, criamos os contextos formais para os caso $1_{(sn,v)}$ e caso $7_{(sn,psP\_sn)}$. A partir desses contextos, construímos as 4 estruturas FCA correspondentes, as quais identificamos como $TourismFCA_{caso1}$, $TourismFCA_{caso7}$, $FinanceFCA_{caso1}$

[2]Optamos por uma frequência baixa na seleção das relações para tentar obter maior quantidade de tuplas. Nosso objetivo foi minimizar a perda de termos relevantes para a tarefa de classificação, ao construir os conceitos formais.

e FinanceFCA$_{caso7}$. Antes de gerarmos as regras a partir dos conceitos dessas estruturas, realizamos uma pequena análise dos seus reticulados.

As Tabelas 9.2 e 9.3 apresentam dados estruturais dos reticulados de conceitos gerados. Usamos nessa avaliação as medidas apresentadas na Seção 8.2.

Analisando-se os dados da Tabela 9.2, pode-se perceber que os textos do domínio Finanças produziram conjuntos de objetos e atributos com quase o dobro de elementos em relação ao domínio de Turismo. Uma das razões pode ser o fato de o conjunto de treino de Finanças possuir cerca de 12% de textos a mais. Outra razão pode estar relacionada à variedade dos textos desses domínios. Em uma análise mais rasa e subjetiva que fizemos, percebemos que os textos do domínio de Turismo são menos abrangentes que os de Finanças. Enquanto os de Turismo abordam temas mais constantes como pontos turísticos, os de Finanças abrangem desde descrições sobre os termos-chave do domínio a aspectos mais políticos, como a biografia de ministros da área de economia.

Tabela 9.2 – Dados das estruturas TourismFCA e FinanceFCA para os casos 1 e 7.

FCA	**#objetos**	**#atributos**	**#conceitos**	**#arestas**	**altura**	**largura**
TourismFCA$_{caso1}$	383	121	239	529	5	106
TourismFCA$_{caso7}$	383	535	343	633	5	218
FinanceFCA$_{caso1}$	631	237	760	2.055	6	360
FinanceFCA$_{caso7}$	631	1018	819	1.919	6	342

Notamos, a partir dos dados da Tabela[3] 9.3, que as estruturas FinanceFCA$_{caso7}$, apesar de superarem em aproximadamente 70% a quantidade de atributos do caso1$_{(sn,v)}$, têm apenas 7% de conceitos a mais e cerca de 7% de arestas a menos. Outro aspecto interessante é que ao analisarmos as medidas de coesão lexical da estrutura FinanceFCA$_{caso7}$, mesmo ainda menores que as da estrutura FinanceFCA$_{caso1}$, seus valores são representativos visto que possuem mais unitários[4]. Esses aspectos podem indicar que os papéis semânticos da estrutura FinanceFCA$_{caso7}$ contribuíram tanto para reduzir o número de grupos (proporcionalmente menos conceitos) quanto para melhorar a qualidade desses grupos (coesão lexical mais expressiva).

Tabela 9.3 – Dados das estruturas TourismFCA e FinanceFCA quanto às medidas CMM e SSM.

Tourism							
FCA	**#unitários(%)**	**CMM$_{TG}$**	**CMM$_{T}$**	**SSM$_{W}$**	**SSM$_{TG}$**	**SSM$_{T}$**	**médiaSSM**
TourismFCA$_{caso1}$	91 (38,1)	28,2	17,8	0,18	0,05	0,02	0,04
TourismFCA$_{caso7}$	190 (49,6)	28,2	17,8	0,09	0,01	0,01	0,04
Finance							
FCA	**#unitários(%)**	**CMM$_{F}$**	**CMM$_{L}$**	**SSM$_{W}$**	**SSM$_{F}$**	**SSM$_{L}$**	**médiaSSM**
FinanceFCA$_{caso1}$	151(19,9)	116,2	91,4	0,33	0,33	0,27	0,31
FinanceFCA$_{caso7}$	283 (44,8)	116,2	91,4	0,20	0,21	0,16	0,19

Após a análise das estruturas FCA, começamos, então, o processo de extração das regras a serem usadas na classificação. Para isso, precisávamos:

[3]Na tabela 9.3, as legendas em subscrito referem-se às bases e ontologias para as quais as medidas CMM e SMM foram calculadas. Desta forma, a abreviação: W significa WordNet; TG, ontologia de Turismo TGPROTON; T, ontologia de Turismo Travel; F, ontologia Finance; e L, ontologia de Finanças LSDIS. Essas ontologias foram comentadas na Seção 5.2.

[4]Cabe lembrar que, "unitários" corresponde aos conceitos cuja cardinalidade do conjunto de objetos formais é igual a 1.

- escolher os conceitos que seriam utilizados para constituir as regras;
- definir como os conceitos selecionados seriam usados para formar as regras e
- determinar quantas regras seriam necessárias para uma boa categorização dos textos.

De acordo com os trabalhos pesquisados, há vários critérios que podem ser usados para selecionar regras [102]. Optamos por um critério simples, mas usual, que é a generalidade dos conceitos [6]. No caso das estruturas FCA, os conceitos mais "gerais" do domínio localizam-se mais próximos às bases dos seus reticulados. São, portanto, aqueles conceitos em que a quantidade de atributos é maior. Desta forma, a exemplo do trabalho de Meddouri e Meddouri [140], primeiramente, organizamos, de forma decrescente, os conceitos de acordo com a cardinalidade dos seus conjuntos de atributos.

Meddouri e Meddouri [140] utilizam apenas os atributos dos conceitos formais para definir as premissas das regras de classificação. Para o objetivo daqueles autores é adequada esta escolha, visto que eles usam as regras para associar instâncias aos conceitos. Para nossa abordagem isso não seria conveniente. Os FCA construídos a partir do caso $1_{(\mathrm{sn,v})}$, por exemplo, gerariam regras cujas premissas seriam formadas apenas por verbos. Levando em conta que queremos categorizar textos e que verbos podem ser aplicados a diferentes domínios, trabalhar apenas verbos deixaria as regras pouco seletivas. Tentamos usar apenas objetos formais nas premissas, mas as regras assim geradas também não resultaram em boas classificações em testes preliminares que realizamos. Optamos, então, por usar tanto objetos quanto atributos formais nas premissas das regras.

Para as regras de estruturas FCA construídas a partir do caso $7_{(\mathrm{sn,psP\_sn})}$, entretanto, consideramos apenas os sintagmas nominais existentes nos atributos. Fizemos isso para evitar que os textos usados nos conjuntos de teste precisassem estar anotados com papéis semânticos para ser classificados. Assim sendo, geramos regras no formato:

IF $objeto_1$ and ... $objeto_i$ and $atributo_1$ and ... $atributo_\mathrm{j}$ THEN categoria

Cabe mencionar que a decisão por usar tanto objetos quanto atributos formais nas premissas das regras teve duas consequências. Uma delas foi que as premissas das regras ficaram muito densas. Por esta razão, usamos apenas um conceito formal na composição de cada regra. A outra consequência foi que essa alta densidade deixou as regras muito específicas.

Como a classificação baseada em regras é realizada, geralmente, a partir do casamento (*matching*) dos termos dos textos com as premissas das regras, regras muito específicas não são interessantes. Uma solução para isso é o particionamento dessas regras em subregras. No entanto, tal solução, além de produzir mais regras, ainda geraria mais um problema que é como agrupar as premissas dessas regras para formar novas regras.

Para evitar esse particionamento, criamos um fator de ativação (fa) para a regra. Esse fator é calculado para cada regra r_i em relação ao texto d_j a ser categorizado, onde $i = 1...q$ e q corresponde ao número total de regras. O fator estabelece uma proporção entre a quantidade de premissas que casam com os termos desse texto com o total de premissas da regra. A categoria c_k com o maior valor de fator de ativação acumulado (fa_{acum}) é definida como a classe do texto d_j. A forma de cálculo desses fatores são apresentadas em 9.1, onde $p(r_i)$ corresponde ao conjunto de premissas da regra r_i e $t(d_j)$, ao conjunto de termos em d_j.

$$fa_{acum}(c_k, d_j) = \sum_{i=1}^{q} fa(r_i, d_j), \text{ onde} \tag{9.1}$$

$$fa(r_i, d_j) = \frac{|p(r_i) \cap t(d_j)|}{|p(r_i)|} \text{ se } c_k \text{ é conclusão de } r_i \text{ e 0 em c.c.}$$

É importante destacar ainda que variamos a quantidade q de regras extraídas de cada estrutura FCA. Analisamos os resultados de categorização para valores q de 23 a 27.

Para avaliar os resultados, usamos medidas usuais em categorização de textos, como precisão (Pr), *recall* (Re) e $F1$ [190]. Calculamos tais medidas por categoria e também de forma geral (macro-médias). A Tabela 9.4 mostra os melhores resultados de categorização obtidos para os conjuntos de teste. Os índices mais altos foram obtidos a partir da extração de 50 regras, 25 regras de cada estrutura FCA. Os dados dessa tabela foram extraídos das tabelas do Apêndice E.

Tabela 9.4 – Melhores resultados da categorização de textos baseada em regras extraídas de estruturas FCA

		FCA		**TourismFCA**			**FinanceFCA**			**Macro-médias**		
conjunto	**q**	**caso**	**rel. I(g,m)**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
$\text{teste}_{\text{Wiki}}$	25	1	(sn,v)	0,64	0,89	0,74	0,82	0,50	0,62	0,73	0,70	0,68
$\text{teste}_{\text{Wiki+PTBS}}$	25	1	(sn,v)	0,58	0,89	0,70	0,85	0,48	0,62	0,71	0,69	0,66
$\text{teste}_{\text{Wiki}}$	25	7	(sn,psP_sn)	0,90	0,94	0,92	0,94	0,90	0,92	**0,92**	**0,92**	**0,92**
$\text{teste}_{\text{Wiki+PTBS}}$	25	7	(sn,psP_sn)	0,68	0,94	0,79	0,93	0,65	0,76	**0,80**	**0,79**	**0,78**

Analisando-se os dados da Tabela 9.4, podemos notar que as regras extraídas a partir de estruturas FCA baseadas no $\text{caso7}_{(\text{sn, psP\_sn})}$ resultam em medidas de avaliação F1 bem superiores às geradas pelo $\text{caso1}_{(\text{sn, v})}$. Cabe mencionar, ainda para essas mesmas estruturas, que os resultados foram melhores para o conjunto de $\text{teste}_{\text{Wiki}}$ possivelmente porque esse conjunto tem a mesma natureza dos textos usados para construir tais estruturas.

9.1.4 Categorização de textos baseada em regras compostas por conceitos ontológicos

Nesse estudo, usamos as ontologias de turismo TGPROTON (O_{TG}) e Travel (O_{T}), e as ontologias de Finanças LSDIS Finance (O_L) e Finance (O_F). A Seção 5.2 descreve brevemente tais ontologias.

Para a escolha dos conceitos, aplicamos o mesmo critério utilizado nas estruturas FCA: generalidade. Buscamos, portanto, conceitos próximos ao topo das hierarquias para compor as premissas das regras. Testamos inicialmente conceitos de nível 2 e, posteriormente, os de nível 3. As premissas de cada regra foram definidas pelo conceito do nível escolhido juntamente com seus superconceitos. Acabamos usando apenas os conceitos de nível 2 em razão da melhor qualidade nos resultados. Para este nível, geramos 28 regras a partir da ontologia TGPROTON, 14 regras a partir da ontologia Travel, 22 regras a partir da ontologia LSDIS Finance e 41 regras a partir da ontologia Finance.

Como tínhamos duas ontologias de cada categoria, testamos primeiramente conjuntos de regras formados a partir das seguintes combinações O_{TG}+O_L, O_{TG}+O_F, O_T+O_L e O_T+O_F. Os resultados de categorização dos conjuntos de teste para essas combinações são apresentados nas Tabelas F.5 e F.6 do Apêndice F. Com o objetivo de melhorar os resultados, escolhemos a combinação de melhor medida F1 para investigar a quantidade de regras mais adequada. Selecionamos, portanto, a combinação O_{TG}+O_F a qual obteve 0,63 e 0,68 em F1 para os conjuntos $\text{teste}_{\text{Wiki}}$ e $\text{teste}_{\text{Wiki+PTBS}}$, respectivamente.

A exemplo dos testes realizados na seção anterior para estruturas FCA, variamos a quantidade q de regras extraídas de cada ontologia. Testamos para q, a princípio, os mesmos valores de 23 a 27. Como os resultados não foram satisfatórios, mudamos os valores de q para o intervalo de 6 a 13. Usando este intervalo para a combinação O_{TG}+O_F, conseguimos uma pequena melhora. Utilizando 18 regras - 9 regras da ontologia TGPROTON e 9 regras da ontologia

Finance - a medida F1 para os teste$_{\text{Wiki}}$ e teste$_{\text{Wiki+PTBS}}$ aumentou, respectivamente, para 0,68 e 0,77 (Tabelas F.7 e F.8 do Apêndice F).

Cabe mencionar que, para selecionar as regras, as organizamos em ordem decrescente conforme a quantidade de premissas. Nosso objetivo era escolher as regras mais abrangentes de cada ontologia. Usamos o mesmo fator de ativação acumulado, detalhado na seção anterior, para determinar a categoria dos textos dos conjuntos de teste. No entanto, tivemos que tornar o fator de ativação fa mais flexível. Dada a presença maior de n-gramas (para $n \geq 2$) como rótulos de classes nas ontologias, tivemos que permitir a ativação das regras também a partir de termos aninhados (*substrings*) desses n-gramas.

Visto que os resultados ainda eram bem inferiores aos obtidos a partir das estruturas FCA, realizamos mais um teste, extraindo regras a partir das ontologias de mesmo domínio conjuntamente ($O_{TG} \cup O_T$, $O_F \cup O_L$). Utilizamos nesses testes o mesmo intervalo para q. Testamos o uso de 6 a 13 regras para cada ontologia. Esses testes são detalhados nas Tabelas F.9 e F.10 do Apêndice F.

A Tabela 9.5 apresenta um resumo dos melhores resultados de categorização obtidos para os conjuntos teste$_{\text{Wiki}}$ e teste$_{\text{Wiki+PTBS}}$. Como podemos observar nessa tabela, os melhores valores para F1 foram obtidos a partir dos últimos testes realizados. Foram necessárias 36 regras (9 de cada ontologia) para atingir os valores 0,71 e 0,77 em F1 para, respectivamente, os conjuntos teste$_{\text{Wiki}}$ e teste$_{\text{Wiki+PTBS}}$. Apesar dos esforços realizados, as regras extraídas a partir dessas ontologias tiveram resultados inferiores aos apresentados pelas regras geradas a partir das estruturas FCA propostas para o conjunto teste$_{\text{Wiki}}$. Para o conjunto teste$_{\text{Wiki+PTBS}}$, no entanto, atingimos medidas F1 equivalentes.

Tabela 9.5 – Melhores resultados para a categorização baseada em regras compostas por conceitos ontológicos

			Tourism			**Finance**			**macro-médias**		
conjunto	**#regras**	**configuração**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
teste$_{\text{Wiki}}$	todas	$O_{TG} + O_F$	0,84	0,38	0,52	0,60	0,93	0,73	0,72	0,66	0,63
teste$_{\text{Wiki+}PTBS}$	todas	$O_{TG} + O_F$	1,0	0,38	0,55	0,67	1,0	0,80	0,84	0,69	0,68
teste$_{\text{Wiki}}$	9	$O_{TG} + O_F$	0,76	0,55	0,64	0,65	0,82	0,72	0,70	0,69	0,68
teste$_{\text{Wiki+}PTBS}$	8	$O_{TG} + O_F$	0,94	0,56	0,70	0,74	0,97	0,84	0,84	0,77	0,77
teste$_{\text{Wiki}}$	9	$O_{TG} \cup O_T + O_F \cup O_L$	0,79	0,78	0,67	0,67	0,85	0,75	**0,73**	**0,72**	**0,71**
teste$_{\text{Wiki+}PTBS}$	9	$O_{TG} \cup O_T + O_F \cup O_L$	0,90	0,59	0,72	0,75	0,95	0,84	**0,83**	**0,77**	**0,78**

Acreditamos que os baixos resultados obtidos nessa abordagem estejam relacionados à relevância semântica das ontologias utilizadas. As ontologias de Turismo, especialmente, possuem um menor detalhamento em conceitos se comparadas às de Finanças. A pouca profundidade das ontologias de Turismo ou mesmo a desproporcional riqueza em conceitos entre as ontologias de domínios distintos podem ter prejudicado os resultados. Além disso, o método simples de extração de conceitos que utilizamos para compor as regras de classificação pode não ter sido o mais adequado.

Na seção seguinte, estudamos a categorização dos textos a partir do algoritmo k-NN.

9.1.5 Categorização de textos baseada no algoritmo k-NN

O processo de classificação implementado pelo algoritmo k-NN consiste em definir a categoria de um documento d_j, não rotulado, que pertence ao conjunto de teste, a partir dos k documentos vizinhos a d_j, que pertencem ao conjunto de treino usado. Os vizinhos são determinados com base em uma métrica que avalia a similaridade entre os termos dos documentos. A categoria mais recorrente entre os documentos vizinhos é associada ao documento d_j [226].

Antes de realizarmos a categorização de textos por meio do algoritmo k-NN, tivemos que pré-processar os textos do conjunto $\text{treino}_{\text{wiki}}$. Como o processador F-EXT-WS já segmenta o texto em *tokens* e os marca com etiquetas POS, esse processo foi relativamente simples. Inicialmente, eliminamos os *tokens* considerados irrelevantes para o processo de categorização de textos, como caracteres especiais e *stopwords*. Sendo que, para descartar estas últimas, usamos as etiquetas POS referentes, por exemplo, a preposições, pronomes, artigos e conjunções. Em seguida, aplicamos o lematizador TreeTagger (Seção 5.3.1) para normalizar os termos restantes. Cabe ressaltar que, não foram usadas as informações sintáticas e semânticas providas pelo anotador F-EXT-WS nesse pré-processamento.

Com o objetivo de definir o conjunto num modelo de *bag-of-words*, usamos, a exemplo de um trabalho que realizamos anteriormente em categorização de textos com o algoritmo k-NN [146] (comentado na Seção 9.2.1.1), seleção por *rank*. Nossa escolha quanto ao tipo de seleção de características a ser utilizado, foi baseada nos bons resultados que encontramos naquele trabalho. Para aplicar a seleção por *rank*, então, contabilizamos a frequência dos termos, resultantes do pré-processamento, em cada categoria. Para cada categoria, escolhemos os n termos de maior ocorrência. Testamos para n os valores 50, 100 e 150. Após a união dos n termos mais relevantes para cada categoria, obtivemos 3 configurações de *bag-of-words*. A seleção por *rank*, para $n = 50$, resultou em um conjunto de 92 termos. Para $n = 100$, encontramos 184 termos significativos. E, por fim, para $n = 150$, testamos uma *bag-of-words* de 276 termos.

A fim de garantir uma uniformidade maior em relação às categorizações realizadas nas seções anteriores, que são baseadas em *matching*, representamos os conjuntos de treino e teste como vetores binários. Usamos, também nos baseando no trabalho descrito em [146], o cosseno como medida de similaridade e três valores de teste para k: 7, 13 e 17.

Os resultados das configurações propostas para o algoritmo k-NN estão no Apêndice F. A partir desses dados, construímos a Tabela 9.6, a qual contém as configurações com os melhores resultados para os arquivos $\text{teste}_{\text{Wiki}}$ e $\text{teste}_{\text{Wiki+PTB}}$. Apesar de o conjunto $\text{teste}_{\text{Wiki+PTB}}$ ter exigido o uso de um número maior de características ($n=150$), o algoritmo k-NN obteve um bom desempenho para ambos os conjuntos. Categorizou os textos com medidas de avaliação F1 de no mínimo 0,92.

Tabela 9.6 – Melhores resultados da categorização de textos baseada no algoritmo k-NN

			Tourism			**Finance**			**Macro-médias**		
conjunto	**n**	**k**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
$\text{teste}_{\text{Wiki}}$	50	17	0,94	0,96	0,95	0,96	0,94	0,95	**0,95**	**0,95**	**0,95**
$\text{teste}_{\text{Wiki+PTBS}}$	150	17	0,89	0,92	0,91	0,94	0,91	0,92	**0,91**	**0,92**	**0,92**

Na seção seguinte, comparamos os resultados de categorização obtidos para as 3 abordagens estudadas.

9.1.6 Comparação dos resultados

Para compararmos as abordagens, reunimos os melhores resultados de categorização na Tabela 9.7. Analisando esta tabela, observamos que as regras extraídas das estruturas FCA geradas a partir de nossa abordagem, que é totalmente automática, obteve resultados iguais ou melhores que aquelas extraídas das ontologias usadas. Claro que devemos considerar que a relevância semântica das ontologias e o fato de não explorarmos profundamente diferentes formas extração de regras a partir delas podem ter influenciado os resultados. Percebemos também, comparando as abordagens baseadas em regras, que o conjunto de $\text{teste}_{\text{Wiki}}$ foi melhor categorizado

pelas regras geradas a partir das estruturas FCA, possivelmente, porque tais estruturas foram construídas a partir dos mesmos *corpora* (WikiFinance e WikiTourism).

Para o algoritmo k-NN, a diferença na natureza dos *corpora* usados nos conjuntos de teste aparentemente não influenciou os resultados obtidos. Comparando os resultados do algoritmo k-NN com os obtidos a partir das estruturas FCA e das ontologias, notamos que para ambos os conjuntos de teste, esse algoritmo foi melhor. As regras extraídas de estruturas FCA, no entanto, para o conjunto $\text{teste}_{\text{Wiki}}$ geraram classificações cujas medidas ficaram muito próximas às do algoritmo k-NN. As regras baseadas em conceitos formais produziram 0,92 em F1 e o algoritmo k-NN, 0,95. Já no caso do conjunto de $\text{teste}_{\text{Wiki+PTBS}}$ tanto as regras baseadas em conceitos formais quanto aquelas baseadas em conceito ontológicos obtiveram F1 de 0,78. Índice bem abaixo do gerado pelo algoritmo k-NN, que foi de 0,92.

Tabela 9.7 – Comparação dos resultados das abordagens usadas para categorização de textos

		Tourism			**Finance**			**Macro-médias**		
conjunto	**abordagem**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
$\text{teste}_{\text{Wiki}}$	k-NN	0,94	0,96	0,95	0,96	0,94	0,95	**0,95**	**0,95**	**0,95**
$\text{teste}_{\text{Wiki+PTBS}}$	k-NN	0,89	0,92	0,91	0,94	0,91	0,92	**0,91**	**0,92**	**0,92**
$\text{teste}_{\text{Wiki}}$	$O_{TG \cup T} + O_{F \cup L}$	0,79	0,78	0,67	0,67	0,85	0,75	0,73	0,72	0,71
$\text{teste}_{\text{Wiki+PTBS}}$	$O_{TG \cup T} + O_{F \cup L}$	0,90	0,59	0,72	0,75	0,95	0,84	**0,83**	0,77	0,78
$\text{teste}_{\text{Wiki}}$	FCA(sn,psP_sn)	0,90	0,94	0,92	0,94	0,90	0,92	**0,92**	**0,92**	**0,92**
$\text{teste}_{\text{Wiki+PTBS}}$	FCA(sn,psP_sn)	0,68	0,94	0,79	0,93	0,65	0,76	**0,80**	0,79	0,78

Embora a construção de estruturas FCA baseadas em papéis semânticos, como as testadas, exija textos com maior riqueza em anotações linguísticas e a geração da própria estrutura demande maior processamento computacional, uma vez definidas e suas regras extraídas, o processo de categorização é mais rápido que o de um algoritmo k-NN. Apesar disso, acreditamos que o mais importante é que conseguimos, por meio dos estudos apresentados, mostrar a aplicabilidade da nossa abordagem.

Cabe mencionarmos que, implementamos em Java o processo de categorização por regras e pelo algoritmo k-NN apresentados nesse estudo.

Na seção seguinte, são comentados alguns estudos que realizamos com textos em português.

9.2 Estudos com *corpus* em Língua Portuguesa

Os estudos apresentados nas seções a seguir foram realizados com o *corpus* PLN-BR CATEG (descrito na Seção 5.1.3). A Seção 9.2.1 mostra os estudos iniciais que fizemos na área de categorização de textos. Levamos em consideração esses estudos ao escolher a tarefa de categorização na avaliação funcional da abordagem proposta na tese. Já a Seção 9.2.2 apresenta os estudos preliminares que realizamos quanto à extração de conceitos e à construção de estruturas conceituais do tipo FCA combinadas com papéis semânticos.

9.2.1 Estudos em categorização de texto

Nesta seção descrevemos brevemente os estudos iniciais que fizemos na área de categorização de textos com o corpus PLN-BR CATEG. Esses estudos geraram publicações conforme citado.

9.2.1.1 Um estudo sobre categorização hierárquica de uma grande coleção de textos em Língua Portuguesa (TIL 2007)

Neste estudo sobre categorização hierárquica de documentos utilizamos 26.606 textos jornalísticos do *corpus* PLN-BR CATEG. Os textos foram lematizados pela ferramenta FORMA [81]

e o processo de categorização foi realizado com o algoritmo k-NN . Implementamos esse algoritmo em linguagem C usando como métrica de similaridade o cosseno. Seguimos a hierarquia de categorias definida por Langie em [119] e testamos duas estratégias de classificação: limiar baseado em *rank* e limiar baseado em relevância. Para representar os documentos, usamos a abordagem *bag-of-words* e definimos os pesos dos termos a partir da medida tf-idf. A seleção dos atributos foi realizada a partir das frequências dos termos nos documentos.

Analisamos neste estudo, de forma experimental, a influência de determinados parâmetros no processo de classificação. Entre os parâmetros analisados estão o número de vizinhos considerados pelo algoritmo (k) e sua relação com a quantidade de documentos usados no treinamento dos classificadores, o número de características escolhidas durante a seleção de atributos, bem como a própria estratégia de classificação.

À época, foi um dos primeiros trabalhos em categorização de textos utilizando um *corpus* de tamanho expressivo em Língua Portuguesa. Mais detalhes podem ser encontrados em [146].

9.2.1.2 *Keywords, k-NN and neural networks: a support for hierarchical categorization of texts in Brazilian Portuguese* (LREC 2008)

Neste estudo também realizamos categorização hierárquica de documentos sobre os mesmos 26.606 textos do *corpus* PLN-BR CATEG (já lematizados pela ferramenta FORMA). Optamos por utilizar, desta vez, palavras-chave na etapa de seleção de atributos. Essa etapa foi realizada com o auxílio da ferramenta Wordsmith Tools (Seção 5.3.6), na qual aplicamos a medida log-likelihood para escolher as palavras conforme as categorias dos documentos. No total, a ferramenta escolheu 500 palavras-chave, as quais foram usadas para representar os documentos do *corpus* (abordagem *bag-of-words*).

Assim como no estudo anterior, seguimos a hierarquia de Langie [119]. No entanto, usamos dois tipos de classificadores: um baseado no algoritmo k-NN (descrito na Seção 9.2.1.1) e outro em redes neurais Multi-Layer Perceptron. Implementamos em linguagem C uma rede neural para cada categoria da hierarquia. Das 28 redes neurais implementadas, 10 foram construídas para o nível 1 da hierarquia e 18, para o nível 2.

Observamos que a categorização dos documentos foi mais precisa com os classificadores neurais. Mais detalhes sobre este estudo podem ser encontrados em [8].

9.2.2 Estudos em extração de conceitos e em estruturas ontológicas

Esta seção apresenta estudos relacionados à extração de conceitos e a estruturas FCA combinadas com papéis semânticos.

9.2.3 Abordagem não supervisionada para extração de conceitos a partir de textos (TIL 2008)

Realizamos um estudo sobre extração de conceitos a partir de textos usando algoritmos de agrupamento. Utilizamos as dependências sintáticas entre os verbos e seus argumentos para identificar os termos e multitermos (n-gramas) relevantes nos textos. A seleção dos termos é realizada através de uma abordagem híbrida que combina as medidas tf-idf e C-Value. Os conceitos são obtidos e organizados através de algoritmos de agrupamento disponíveis na ferramenta CLUTO (Seção 5.3.6). Nossos experimentos foram realizados em 4.407 documentos da seção Esportes do *corpus* PLN-BR CATEG, e a qualidade semântica dos *clusters* foi avaliada manualmente.

A abordagem híbrida, bem como o método de seleção de termos propostos mostraram-se adequados na escolha de termos relevantes para a identificação de conceitos. Apesar de relatarmos vários problemas na geração dos grupos de conceitos, os resultados mostram a

viabilidade da metodologia. Mais detalhes sobre esse trabalho podem ser encontrados em [145].

9.2.4 Estruturas FCA e papéis semânticos

Esta seção descreve brevemente um estudo que realizamos para a Língua Portuguesa, voltado à construção de estruturas conceituais baseadas em FCA e em papéis semânticos. Usamos nesse estudo preliminar cerca de 58% dos textos da seção Esportes do *corpus* PLNBR-CATEG.

Para extrair as relações verbo-argumento dos textos, já lematizados pela ferramenta FORMA, desenvolvemos, em linguagem C, um *parser* baseado em expressões regulares. Verbos de ligação e auxiliares como "ser" e "ter" não foram considerados. Foram extraídos, como argumentos dos verbos, os sintagmas nominais que desempenhavam a função sintática de objeto direto ou indireto. Usamos uma heurística simples para eliminar as entidades nomeadas: excluímos todos os sintagmas que iniciavam com letra maiúscula e não estavam no início das sentenças. Tivemos também que criar uma pequena *stoplist* para eliminar termos referentes a datas, períodos e dias da semana pois, como o *corpus* era jornalístico, esses termos eram muito frequentes.

Para permitir uma análise qualitativa, utilizamos um exemplo de estrutura FCA. Essa estrutura foi gerada a partir das sementes "jogo" e "campeonato". Seu contexto formal se baseou em relações verbo-argumento cuja frequência em documentos fosse superior a 9. Para estreitar ainda mais a relação semântica entre as selecionadas, foram descartadas aquelas cujos verbos não estivessem relacionados a pelo menos 3 diferentes sintagmas nominais (argumentos).

Em seguida, traduzimos para o inglês as sentenças cujas relações foram selecionadas. Para realizar a anotação semântica usamos um dicionário de verbos para Língua Portuguesa [24], um etiquetador de papéis semânticos para a Língua Inglesa [174] e a VerbNet. Cabe ressaltar que tivemos que usar um etiquetador de papéis semânticos para a Língua Inglesa devido à inexistência desse recurso para Língua Portuguesa. O etiquetador usado Illinois SRL disponibilizado pelo Cognitive Computation Group e comentado na Seção 5.3.4. Esse etiquetador segue a anotação do PropBank para identificar os papéis semânticos. Como essa anotação não é padrão para todos os verbos e nem corresponde aos nomes geralmente usados na literatura para identificar papéis (Agent, Patient, ...), usamos o mapeamento referente a papéis semânticos entre o PropBank e a VerbNet, descrito em [129], para etiquetar as relações. As etiquetas assim definidas são verificadas com o auxílio do dicionário de verbos e, então, usadas como atributos durante a geração dos conceitos que formam o FCA.

Após a anotação analisamos para cada contexto formal, pelo menos 6 sentenças (escolhidas aleatoriamente) nas quais as relações apareciam. Isso permitiu a identificação de alguns erros do *parser* e do etiquetador Illinois SRL.

Para formar o contexto formal da estrutura FCA de nosso exemplo, usamos os sintagmas nominais (argumentos dos verbos) como objetos e elementos da forma *verb(semanticRole)* como atributos formais. Essa estrutura-exemplo é apresentada na Figura 9.2.

Analisando a estrutura FCA gerada a partir das sementes "jogo" (*game*) e "campeonato" (*championship*), conseguimos distinções interessantes de significado, principalmente para os papéis Agent, Manner e Location.

Os argumentos do verbo "dizer" (*to say*), por exemplo, foram separados corretamente em Agent e Topic (assunto de uma comunicação). Todos os objetos anotados como Agent, nesse caso, eram denominações dadas a pessoas conforme a função que executam (atacante, jogador, treinador, ...). Em estruturas conceituais, como ontologias, os objetos Agent poderiam ser indicados, em uma abordagem semiautomática, como subclasses de *Pessoa*. Poderiam, também, ser usados para encontrar as instâncias de *Pessoa* no texto.

Já o objeto *"casa" (home)* foi etiquetado como Location no contexto de "jogar" (*to play*). Todos os demais objetos de "jogar" foram anotados como Theme. Confirmamos que os eti-

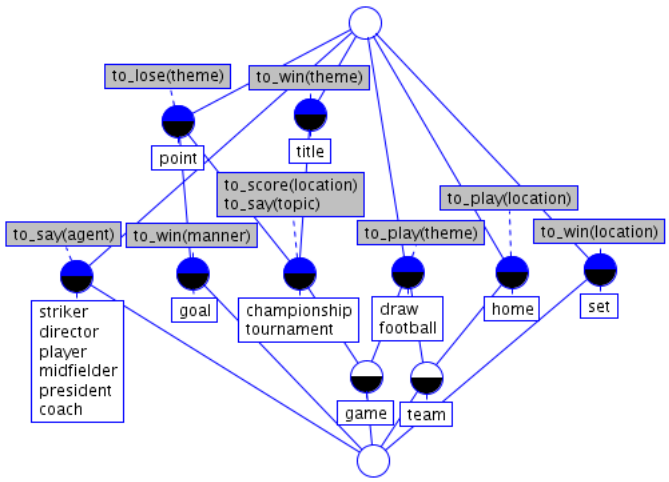


Figura 9.2 – Estrutura FCA para as sementes "jogo" e "campeonato".

quetadores semânticos são ainda muito dependentes da sintaxe, especialmente no que se refere ao papel Location. Os objetos "partida" (*set*) e "campeonato" (*championship*), embora não sejam locais concretos (físicos), foram rotulados como Location para os verbos "vencer" (*to win*) e "pontuar" (*to score*), respectivamente. O uso desse papel para construção de estruturas conceituais também é interessante, no entanto, deve-se levar em consideração que ele indica o "contexto" em que o evento ocorreu e isso nem sempre corresponde a um lugar concreto.

Outra diferenciação interessante foi a atribuída ao objeto "gol" (*goal*). Ele foi anotado como uma forma (Manner) de vencer. Isso aconteceu em sentenças como: *"O técnico Paulo César Carpegiani, que espera vencer por dois ou mais gols de diferença para ..."*

Um aspecto também interessante dos papéis semânticos diz respeito aos sinônimos. Sintagmas nominais (objetos) que compartilham verbos e são etiquetados para esses verbos com os mesmos papéis semânticos, podem ser candidatos a sinônimos, como é o caso de "campeonato" e "torneio".

O papel mais frequente foi Theme. Aparentemente, todas as palavras etiquetadas com esse papel eram de fato relevantes para o domínio. Como acreditamos que a construção de estruturas conceituais seja um processo interativo e iterativo, podemos usar os objetos Theme como as sementes da próxima iteração, o que permitiria refinar passo a passo os conceitos da estrutura.

Para finalizar, é importante destacar que, mesmo com a falta de recursos, fizemos estudos iniciais de nossa proposta em Língua Portuguesa.

9.3 Considerações sobre este capítulo

Como já mencionado, os estudos em categorização de texto para Língua Inglesa mostram que nossa proposta de utilizar estruturas FCA baseadas em papéis semânticos é aplicável. É importante ressaltar que mesmo diante de diferentes limitações, as estruturas FCA construídas segundo nossa abordagem conseguiram gerar boas regras de classificação.

Observamos que durante a realização dos testes com os *corpora* WikiFinance e WikiTourism, o *parser* que implementamos para eliminar *tags* e elementos indesejáveis em tais textos não foi tão eficiente. Ainda restaram marcadores de itens, por exemplo, em alguns textos. Essa limpeza deficiente, sob alguns aspectos, provavelmente prejudicou o desempenho do processador F-EXT-WS quanto à anotação semântica. Notamos também que a simplicidade do *parser* que desenvolvemos para extrair sintagmas nominais dos textos, produziu, em alguns casos, unigramas menos significativos. Se associarmos essas limitações à precisão possivelmente mais baixa de anotadores semânticos, como o utilizado, para domínios diferentes do Finanças, podemos

afirmar que os resultados obtidos são satisfatórios.

Nesse capítulo ainda apresentamos estudos relacionados à Língua Portuguesa. Realizamos estudos em extração de conceitos e categorização de textos. Desenvolvemos também um estudo preliminar usando estruturas FCA com papéis semânticos, convivendo com a ausência de anotadores dessa natureza específicos para o português. Embora a tradução das sentenças para a Língua Inglesa tenha dificultado esse estudo, consideramos os resultados promissores.

10. CONCLUSÕES E METODOLOGIAS

10.1 Considerações gerais

Neste trabalho, estudamos abordagens tradicionais para aprendizagem de estruturas ontológicas a partir de textos e propomos a construção de tais estruturas a partir do método FCA e de papéis semânticos.

Apesar de termos encontrado, ao longo de nossa pesquisa, trabalhos que resultaram em estruturas conceituais relevantes, percebemos que não há consenso, entre os pesquisadores, no que se refere a diferentes aspectos referentes à aprendizagem de tais estruturas a partir de textos. Há divergência quanto ao conceito de ontologia e divergência quanto a quais estruturas conceituais podem, de fato, ser assim chamadas. Muitas estruturas ditas ontologias são apenas taxonomias e não incluem axiomas e nem relações transversais. Essa discordância entre os pesquisadores foi uma das razões, inclusive, pelas quais optamos pelo uso de termos como "estruturas ontológicas" e "estruturas conceituais", para nos referirmos aos reticulados de conceitos gerados pelo método FCA.

Quanto à metodologia, percebe-se algumas linhas gerais de procedimento, como as tarefas enumeradas por Cimiano (Seção 2.5.1) para aprendizagem dessas estruturas a partir de textos. Notamos uma certa preferência, por parte dos pesquisadores, por abordagens híbridas e pelo uso de recursos *web* na implementação da maioria dessas tarefas. Muitos trabalhos combinam técnicas linguísticas, estatísticas e de aprendizagem de máquina para realizá-las, sendo uma prática comum o uso de *corpora* construídos a partir de documentos *web* e de textos da Wikipédia. Observamos também o uso recorrente da WordNet como um recurso de apoio à identificação de conceitos e relações semânticas.

No entanto, percebemos que tarefas referentes à identificação e à etiquetagem de relações transversais, bem como tarefas responsáveis pela definição de axiomas, nem sempre estavam presentes, nos trabalhos pesquisados. E é justamente em tais tarefas que os papéis semânticos têm sido usados com maior intensidade. Em trabalhos mais recentes [11, 22, 193], eles são utilizados para identificar e rotular relações transversais, bem como para prover restrições semânticas úteis na definição de axiomas.

Embora papéis semânticos sejam um recurso interessante para extração e representação de estruturas ontológicas, há ao menos duas dificuldades importantes no que tange a sua utilização: a falta de uma lista consensual de papéis semânticos e a imprecisão dos anotadores atuais para domínios que não sejam o do PropBank. Embora as etiquetas numéricas usadas na anotação PropBank tenham, de certa forma, contornado o problema referente à identificação dos papéis, a dificuldade no seu uso ainda persiste pois, em aplicações que envolvem representação de conhecimento, é necessário determinar o significado de cada etiqueta. Além disso, mesmo no caso da Língua Inglesa, para a qual os avanços na área são maiores, há poucos *corpora* anotados semanticamente que possam ser usados no treinamento de etiquetadores semânticos. Essa escassez tem impactado no aperfeiçoamento dessas ferramentas.

Agregadas a essas dificuldades de anotação semântica, estão ainda aquelas inerentes a qualquer abordagem que busque informações a partir de textos. Essas dificuldades, somadas, acabam tornando essencial o olhar humano. Desta forma, é muito comum o uso de abordagens semiautomáticas na construção de estruturas ontológicas a partir de textos. Isso, inclusive, se estende ao processo avaliativo de tais estruturas, comumente conduzido sob a forma de avaliação humana.

Avaliações intrínsecas de cunho puramente semântico, com frequência são realizadas manualmente. Quando o objetivo é analisar a estrutura ontológica enquanto estrutura de representação

de conhecimento, a avaliação geralmente é realizada por especialistas humanos. A ausência de métodos e métricas formais de uso amplo, além de prejudicar as avaliações em si, também tem inviabilizado a comparação dos resultados obtidos em diferentes pesquisas. Mesmo em avaliações extrínsecas, de caráter funcional, como as que utilizam ontologias de referência [41] ou comparam os resultados no contexto de uma aplicação [152], se ressentem dessa ausência.

Do ponto de vista tecnológico, não há ainda ferramentas que viabilizem a avaliação automática com componente semântico das estruturas conceituais geradas a partir de textos. A falta de *benchmarks* para essa área, por exemplo, inviabiliza a comparação de metodologias. Desta forma, torna-se difícil decidir que método e técnica são mais adequados para a realização de uma tarefa em *corpora* de domínio e de natureza distintas [80].

Todas essas lacunas mostram que a área de aprendizagem de estruturas ontológicas a partir de textos ainda tem muito espaço de investigação. E é dentro desse espaço que o método FCA tem despertado o interesse dos pesquisadores e, consequentemente, vem se destacando como uma alternativa enquanto método de agrupamento conceitual na construção de estruturas ontológicas. Em ciência da computação, especialmente, o interesse por FCA e suas extensões se justifica pelo fato de o método ser adequado para a análise de dados e ser promissor como forma de representação de conhecimento. Suas inerentes características de agrupar, relacionar e organizar os dados de forma hierárquica [167, 231] e de, ainda, prover descrições intensionais que facilitam a interpretação dos grupos gerados [98], o tornam um método de agrupamento conceitual muito interessante.

Em PLN, ele se destaca por permitir diferentes representações conceituais que refletem as diversas formas como os dados aparecem relacionados nos textos. Sendo indicado, portanto, para análises linguísticas, pois gera estruturas que possibilitam o estudo de relacionamentos sintáticos e semânticos, inclusive para desambiguação de sentido [167].

A principal desvantagem do método, no entanto, é de ordem computacional. À medida que a quantidade de dados e de relacionamentos entre esses dados aumentam, cresce de forma exponencial a complexidade de geração dos reticulados de conceitos provenientes da aplicação do método. Esse aspecto, entretanto, é atenuado pelo fato de utilizarmos informações provenientes de textos. O problema da esparsidade das relações encontradas no textos, tal como as do tipo verbo-argumento, ao mesmo tempo que reduz a riqueza da informação extraída, diminui também a complexidade de construção dos reticulados de conceitos. Embora essa esparsidade não seja desejável no contexto da extração de informações, pesquisadores que utilizam o método para relacionar informações textuais têm relatado complexidade próxima à linear [40].

Apesar do problema referente à complexidade computacional, os resultados obtidos pelos pesquisadores com o FCA, enquanto método de agrupamento conceitual, têm sido motivadores. Este fato, aliado à aplicabilidade dos papéis semânticos na identificação de relações não taxonômicas e o recente surgimento de etiquetadores automáticos para tais papéis nos levaram a indagar se relações semânticas dessa natureza não poderiam melhorar a qualidade dos conceitos formais gerados pelo método FCA. Como, ao longo de nossa pesquisa, não encontramos trabalhos atuais que, juntamente com o método FCA, explorassem aspectos semânticos, mais especificamente classes de verbos e papéis semânticos, direcionamos nosso estudo nesse sentido.

Dado que os etiquetadores automáticos de papéis semânticos estão disponíveis para a Língua Inglesa, concentramos nossa investigação em *corpora* nessa língua. Mesmo com a carência desse recurso para a Língua Portuguesa, realizamos estudos preliminares em um *corpus* de textos em português: PLN-BR CATEG.

Usamos em Língua Inglesa os *corpora* Penn TreeBank Sample, SemLink 1.1 e Wikicorpus 1.0 sendo que, no caso desse último, trabalhamos apenas com um subconjunto dos textos. Geramos, a partir do Wikicorpus 1.0, pequenos *corpora* dos domínios de Finanças e Turismo, aos quais chamamos de WikiFinance e WikiTourism. Enquanto que o SemLink 1.1 já continha as anotações semânticas referentes ao Penn TreeBank Sample e eram adequadas ao nosso estudo,

os *corpora* WikiFinance e WikiTourism não dispunham de tal informação. Tivemos então que anotá-los e para isso usamos o processador F-EXT-WS. Cabe lembrar que o *corpus* SemLink 1.1 é uma extensão do PropBank. Nessa extensão, foram incluídas informações quanto às classes VerbNet dos verbos e foram mapeados os papéis semânticos VerbNet às etiquetas numéricas PropBank correspondentes.

Na investigação que fizemos com o uso dos *corpora* Penn TreeBank Sample e SemLink 1.1 conseguimos estudar os papéis semânticos a partir de rótulos mais tradicionais, tais como Agent e Patient. No caso dos *corpora* WikiFinance e WikiTourism, anotados semanticamente pelo processador F-EXT-WS, tivemos que conviver com rótulos numéricos ao estilo PropBank para tais papéis. Essa diferença nos rótulos nos obrigou a tratar tais anotações de forma diferenciada.

Independente disso enfrentamos, no pré-processamento de todos esses *corpora,* dificuldades quanto à extração dos sintagmas nominais anotados com papéis semânticos. Como a tarefa de extração de sintagmas nominais é relativamente complexa, tivemos que criar um conjunto de heurísticas para realizá-la. Embora esse tratamento simplificado tenha provocado consequências, visto que foram gerados menos sintagmas do que de fato existiam nos textos, ou alguns de menor relevância, conseguimos mesmo assim realizar nossa investigação de forma satisfatória.

Um dos aspectos fundamentais de nossa pesquisa foi a inclusão das informações semânticas em estruturas do tipo FCA. Inicialmente, estudamos o método RCA, que é uma extensão do FCA. Como esse método é indicado, na literatura, como o mais adequado para representar explicitamente relações não taxonômicas, tais como as relações semânticas definidas pelos papéis, analisamos a viabilidade de seu uso em nossa abordagem. A generalização das relações entre objetos formais para relações entre conceitos formais, na análise preliminar que realizamos, produziu relações inicialmente inexistentes entre os objetos. A estrutura RCA assim gerada não explicitava tais ausências claramente e imaginamos que isso poderia levar a erros de interpretação quanto aos conceitos gerados. Além disso, observamos a falta de ferramentas para geração de estruturas RCA. Esses fatores nos levaram a descartar tal método.

O uso do *corpus* SemLink 1.1 também nos permitiu analisar algumas classes VerbNet quanto aos papéis semânticos frequentemente associados aos verbos dessas classes. Observamos que tais classes, de fato, delimitavam os papéis semânticos. Por esta razão, acreditamos que elas pudessem ser um caminho de pesquisa interessante para a investigação de relações não taxonômicas mediadas por papéis semânticos. Contudo, como a VerbNet não é completa, seria necessário estudar métodos e heurísticas capazes de classificar os verbos ausentes nessa base. Com esses mecanismos de classificação e com o auxílio dos etiquetadores de papéis semânticos seria possível determinarmos, de forma mais objetiva, as relações entre papéis e sua relevância para diferentes domínios. Acreditamos que, para esse tipo de análise, o mais indicado seria o uso de *corpora* de natureza conceitual. Percebemos, como já mencionado, que a natureza do *corpus* influencia na frequência de determinados papéis semânticos, podendo "reduzir a importância" daqueles cuja relevância pode ser igual ou maior para o domínio.

Ainda no que se refere às relações não taxonômicas, acreditamos que as classes de verbos possam ajudar a definir rótulos de caráter mais geral para tais relações. Isso exigiria, entretanto, associar às denominações numéricas atuais, rótulos textuais. Embora essa ideia pareça atraente, acreditamos ser difícil sua realização de forma automática. Para que os rótulos façam sentido, têm que ser definidos de acordo com o domínio e, nesse caso, a sugestão ou mesmo crítica humana seriam essenciais. Por outro lado, como analisamos apenas duas classes VerbNet, talvez a complexidade quanto à definição desses rótulos textuais seja maior do que tenhamos conhecimento. É possível que a relação estabelecida entre os papéis de uma mesma classe não seja tão clara. De qualquer forma, acreditamos que caiba investigar tal possibilidade.

Voltando à questão referente à inclusão de papéis semânticos em estruturas FCA, para determinar a forma como isso seria realizado, tivemos que analisar diferentes configurações de contextos formais. Um dos problemas que enfrentamos foi a definição do procedimento

avaliativo que seria utilizado nessa fase. Precisávamos analisar os conceitos formais do ponto de vista estrutural e também semântico. Só assim poderíamos medir a contribuição de classes de verbos e papéis semânticos na construção desses conceitos.

Dada a escassez de medidas estruturais adequadas a nosso propósito, principalmente, no que se refere a conceitos formais, propusemos a adaptação da medida SSM. Escolhemos tal medida estrutural por ela levar em conta aspectos semânticos. Para que ela atendesse ao nosso fim, operacionalizamos sua aplicação a conceitos. Desta forma, ela passou a funcionar como uma medida de coesão lexical. Embora tal medida, se aplicada de forma isolada, não seja conclusiva, aliada a outras medidas estruturais a mesma nos forneceu dados que ajudaram a analisar as configurações propostas e a escolher as mais promissoras, no contexto de nossa pesquisa.

Como o auxílio dessa medida, notamos que as classes de verbos, além de aumentarem a coesão lexical, ajudaram a reduzir a complexidade de construção do reticulado FCA, na medida em que geraram menos conceitos e arestas. Já os papéis semânticos mostraram-se mais efetivos, no aspecto coesão, principalmente quando eram usados como atributos nos contextos formais estudados. Apesar desses resultados, do ponto de vista intensional e simbólico, o uso, como atributos formais, de rótulos numéricos como os que identificam as classes de verbos é menos informativo do que o uso de verbos. Entendemos, igualmente, como menos expressivos sob o aspecto intensional, os atributos contendo papéis semânticos dissociados de seus contextos. Em um contexto, o papel semântico do sintagma nominal assim anotado é mais evidente e, consequentemente, mais relevante, pois as relações com outros elementos do domínio delimitam o seu significado.

Por esta razão acabamos convergindo para uma representação em que os pares objeto-atributo expressavam relações entre papéis semânticos. Nessa representação, os atributos eram formados por contextos lexicossemânticos. Os atributos assim definidos nos pareceram intensionalmente mais descritivos, ainda que inicialmente (antes da aplicação das heurísticas descritas na Seção 8.4) tal configuração tenha produzido conceitos menos coesos. Com base nos estudos que realizamos para definir tal representação, descrevemos, na forma de um procedimento, na Seção 10.2, a metodologia que usamos para construir estruturas FCA a partir de tais contextos lexicossemânticos. Chamamos esta metodologia de Semantic FCA (SFCA) e a consideramos um dos principais resultados de nossa investigação.

Com o objetivo de complementar ainda mais nossos estudos e também analisar a viabilidade quanto à aplicação da proposta, avaliamos a contribuição das informações semânticas na construção de conceitos formais com base nos *corpora* WikiFinance e WikiTourism. Nesse caso, contudo, optamos por medidas de avaliação de ordem funcional no contexto da tarefa de categorização de textos.

Escolhemos trabalhar, nessa etapa da investigação, com a classificação de textos baseada em regras, embasados em trabalhos relacionados. Tal forma de trabalho nos permitiu comparar resultados provenientes das regras extraídas das estruturas FCA enriquecidas semanticamente (construídas conforme nossa abordagem) daquelas extraídas de ontologias. Em todos os testes realizados, as regras extraídas a partir das estruturas FCA geraram resultados de classificação iguais ou superiores às produzidas pelas obtidas a partir das ontologias. Cabe ressaltar que há ainda espaço de investigação quanto à extração de regras a partir dessa estruturas, visto que o método usado e a relevância semântica das ontologias usadas no processo comparativo podem ter influenciado de forma significativa os resultados.

Comparamos, também, os resultados obtidos a partir de nossa abordagem aos de um classificador k-NN. Em alguns casos conseguimos resultados muito próximos aos desse classificador. Obviamente que o esforço de anotação linguística para o uso de um classificador desse tipo é muito menor do que o necessário para gerar uma estrutura FCA. Além disso, a complexidade envolvida na construção do reticulado de conceitos também aumenta a demanda computacional. No entanto, uma vez construída essa estrutura conceitual e extraídas as suas regras, o

processo de categorização é mais rápido do que o de um classificador k-NN.

Embora nem sempre nossa abordagem tenha gerado resultados superiores aos do classificador k-NN, consideramos os resultados obtidos como satisfatórios, visto que, mesmo diante de diferentes limitações, as estruturas FCA enriquecidas semanticamente conseguiram gerar boas regras de classificação. Embora o pré-processamento dos *corpora* WikiFinance e WikiTourism, em alguns casos, tenha comprometido o desempenho do anotador F-EXT-WS, e a simplicidade das heurísticas usadas na extração dos sintagmas nominais dos textos tenha sido menos efetiva, nossa abordagem mostrou-se aplicável. Assim, também consideramos esse classificador outra contribuição de nossa pesquisa. Na Seção 10.3, descrevemos a metodologia que permite construir e utilizar um classificador de textos baseado em estruturas SFCA.

Quanto aos estudos realizados para a Língua Portuguesa, a ausência de recursos linguísticos é um fator limitante. Com base nos estudos preliminares que realizamos, acreditamos que nossa abordagem também seja viável para o português. No entanto sua aplicação exigirá um esforço manual, dada a falta de anotadores de papéis semânticos para essa língua.

Cabe mencionar ainda que, quanto às formas de seleção estudadas na Seção 8.3, o uso de sementes em conjunto com o "operador mais" mostrou-se um caminho apropriado para gerar estruturas FCA. Essa abordagem, além de melhorar a coesão lexical dos conceitos formais, também reduziu a complexidade de construção do reticulado FCA.

10.2 SFCA: metodologia para construção de estruturas FCA baseadas em papéis semânticos

Nesta seção descrevemos a metodologia Semantic FCA (SFCA) que permite construir uma estrutura FCA enriquecida com papéis semânticos. Para aplicar tal metodologia, o *corpus* a partir do qual a estrutura será gerada deve ser anotado com informações lexicossemânticas. Para isso, o *corpus* deve ser submetido a um etiquetador de papéis semânticos e a um anotador de POS. Desta forma, será possível identificar os verbos, os seus argumentos e os papéis semânticos desses argumentos.

Com o *corpus* assim anotado, a construção de uma estrutura SFCA consiste em:

1. Normalizar morfologicamente os termos das sentenças do *corpus* por meio de um lematizador.

2. Analisar as sentenças, identificando e extraindo seus verbos, os argumentos desses verbos e os papéis semânticos associados a esses argumentos. Desconsiderar sentenças ou trechos de sentenças cujos argumentos de verbos não tenham sido anotados semanticamente.

3. Identificar os sintagmas nominais existentes nos argumentos anotados com papéis semânticos, e considerar apenas aqueles formados por substantivos comuns. Para efetuar esse passo, será necessário criar heurísticas para tratar os sintagmas nominais, tais como as descritas na Seção 6.4.2.

4. Formar tuplas, usando as informações extraídas das sentenças nos passos 2 e 3, no seguinte formato: (sn_1, ps_1, sn_2, ps_2), onde sn_i e ps_i correspondem, respectivamente, ao sintagma nominal e ao seu papel semântico. Nas tuplas, os sintagmas nominais devem ser constituídos pelos lemas de seus substantivos. Cada tupla deve conter apenas dois argumentos (sintagmas nominais) cujos papéis semânticos foram atribuídos por um verbo, em uma mesma sentença. Se, na sentença, o verbo tiver mais de dois argumentos anotados semanticamente, devem ser formadas tantas tuplas quantas forem as combinações desses argumentos dois a dois. Por outro lado, se, na sentença, o verbo tiver apenas um argumento anotado com papel semântico, tal informação será insuficiente para formar uma tupla e, portanto, deve ser descartada.

5. Construir, a partir das tuplas formadas no passo 4, os pares $(objeto, atributo)$ da seguinte forma: $(sn_1, ps_1\_sn_2)$ e $(sn_2, ps_2\_sn_1)$. Incluir esses pares em uma lista, contabilizando as suas frequências absolutas.

6. Selecionar os pares $(objeto, atributo)$ mais significativos, no caso de uma lista de pares $(objeto, atributo)$ muito longa. A frequência absoluta pode ser usada para escolher os mais representativos.

7. Aplicar alguma técnica de agrupamento para os atributos quando esses forem muito específicos, a fim de evitar um contexto formal muito esparso. Para esse fim, pode ser usado o coeficiente Dice tal como fizemos em nossa investigação (Seção 8.4).

8. Construir o contexto formal usando os pares $(objeto, atributo)$ resultantes dos passos 5, 6 e 7.

9. Gerar a estrutura SFCA a partir do contexto formal construído no passo anterior. O reticulado de conceitos pode ser construído a partir de algoritmos como os comentados na Seção 3.5.2 ou a partir de uma ferramenta com esse propósito, tal como as mencionadas na Seção 5.3.5.

10.3 Metodologia para categorização de documentos baseada em estruturas SFCA

Nesta seção descrevemos a metodologia utilizada para construir e utilizar um classificador baseado em regras, em que tais regras são formadas por conceitos formais extraídos de estruturas SFCA. Como em todo processo de categorização, é necessário, inicialmente, separar os textos por categoria e particioná-los proporcionalmente em conjuntos de treino e de teste. Para particionar os textos em tais conjuntos pode ser usada a abordagem *train-and-test* [190], que usamos em nossa pesquisa.

Uma vez definidos os conjuntos de treino e teste por categoria, pode-se aplicar a metodologia de categorização de textos baseada em estruturas SFCA, a qual se divide em duas etapas detalhadas a seguir.

- Etapa 1: corresponde à geração das regras do classificador. Para defini-las, é necessário:

    1. Construir uma estrutura SFCA para cada categoria a partir do respectivo conjunto de treino. Cada estrutura SFCA deve ser gerada conforme descrito na seção anterior.
    2. Extrair de cada estrutura SFCA construída, os q conceitos formais mais gerais. Nesses conceitos a quantidade de atributos, usualmente, é maior e eles encontram-se próximos à base dos seus reticulados SFCA. Uma forma de identificar os conceitos mais gerais de uma estrutura SFCA é ordená-los de forma decrescente, conforme a cardinalidade dos seus atributos. Após a ordenação dos conceitos dos SFCA, basta escolher os q primeiros de cada estrutura para obtermos os conceitos mais gerais de cada categoria.
    3. Definir as regras do classificador, a partir dos conceitos gerais de cada categoria extraídos no passo anterior. Cada conceito formal gera uma regra do classificador. As regras devem possuir o formato descrito a seguir, no qual os objetos e atributos dos conceitos formais constituem as premissas e as categorias, as conclusões. Cabe ressaltar que os atributos que compõem as premissas devem conter apenas sintagmas nominais. A exclusão dos papéis semânticos teve como propósito evitar que, durante a classificação, o conjunto de teste tenha que ser anotado com tais informações semânticas.

IF $objeto_1$ and ... $objeto_i$ and $atributo_1$ and ... $atributo_j$ THEN categoria

- Etapa 2: refere-se à categorização dos documentos propriamente dita. Tal categorização será realizada com base nas regras do classificador construído na Etapa 1. Desta forma, para classificar um documento d_j do conjunto de teste é preciso:

  1. Extrair os sintagmas nominais do documento d_j, formando o conjunto $t(d_j)$.
  2. Calcular o fator de ativação (fa) de d_j para cada regra r_i do classificador. Este fator é calculado conforme apresentado na Equação 10.1, onde $p(r_i)$ corresponde ao conjunto de premissas de r_i. Tal fator estabelece uma proporção entre a quantidade de premissas que casam com $t(d_j)$ com o total de premissas da referida regra.

$$fa(r_i, d_j) = \frac{|p(r_i) \cap t(d_j)|}{|p(r_i)|} \tag{10.1}$$

  3. Calcular, para cada categoria c_k, o fator de ativação acumulado (fa_{acum}) para o documento d_j. Esse fator é calculado conforme apresentado na Equação 10.2, onde q corresponde ao número de regras do classificador em que a conclusão é c_k. A categoria c_k com o maior valor fa_{acum} é definida como a classe do texto d_j.

$$fa_{acum}(c_k, d_j) = \sum_{i=1}^{q} fa(r_i, d_j) \tag{10.2}$$

Empregamos medidas usuais em categorização de textos, como precisão (Pr), *recall* (Re) e $F1$ [190] para avaliar os resultados do classificador baseado em SFCA. A qualidade desse classificador está relacionada ao seu conjunto de regras. Para obter um bom conjunto de regras, é interessante realizar testes variando a quantidade q de conceitos formais extraídos das estruturas SFCA. É aconselhável utilizar o valor q que gerar a melhor medida $F1$ para o conjunto de teste usado na etapa de categorização.

10.4 Trabalhos futuros

Nessa seção, indicamos estudos que podem melhorar a abordagem apresentada, quanto ao enriquecimento de estruturas FCA com informações semânticas. São eles:

- Estudo mais aprofundado de estruturas RCA: embora tenhamos descartado o uso da estrutura RCA em nossa pesquisa, acreditamos que tal extensão deva ser estudada em maior profundidade. Possivelmente, a inclusão de mais relações tais como do tipo substantivo-adjetivo ou mesmo substantivo-advérbio contribua na qualificação dos conceitos formais. Essa qualificação pode tornar mais viável a generalização de relações entre objetos para relações entre conceitos, proposta pelo método RCA, para construção de estruturas conceituais.

- Construção de um *parser* mais eficiente para extração de textos Wikipédia em formato *raw*: percebemos que limitações do *parser* implementado para este fim trouxeram dificuldades ao processador F-EXT-WS quanto à anotação de papéis semânticos.

- Construção ou utilização de um *parser* mais robusto para extração de sintagmas nominais: a simplicidade das heurísticas, em alguns casos, resultou em menor quantidade de sintagmas e em sintagmas menos expressivos.

- Estudo de heurísticas e métodos para identificação de classes de verbos: de acordo com nosso estudo, as classes ajudam a construir conceitos formais mais coesos. No entanto, não conseguimos analisar essa contribuição para estruturas FCA a partir de *corpora* não anotados com tal informação, devido à incompletude da VerbNet. A implementação dessas heurísticas e métodos viabilizaria também um estudo mais aprofundado das relações transversais mediadas por papéis semânticos, e sua relevância em diferentes domínios.

- Investigação de outros métodos de extração de regras a partir das estruturas FCA e das ontologias usadas no processo comparativo: tanto o método de extração e aplicação das regras quanto a relevância das ontologias utilizadas na investigação realizada, influenciaram os resultados obtidos. Um estudo mais profundo nesse sentido se faz necessário. Assim como é realizado por algoritmos usados em regras de associação, poderíamos definir fatores de suporte e confiança para selecionar os conceitos e as regras durante o processo de classificação.

- Implementação de um programa que permita exportar as estruturas FCA para o formato OWL: isso potencializaria a utilidade da proposta, visto que tornaria mais fácil o seu uso para a criação de ontologias. Cabe mencionar que, nesse caso, seria necessário primeiramente converter a estrutura FCA para uma hierarquia de conceitos, antes de exportá-la para OWL. O procedimento de transformação de uma estrutura FCA em uma hierarquia é descrito por Cimiano em [41].

- Implementação de um programa que permita a navegação dinâmica nos nodos de uma estrutura FCA, a fim de facilitar a sua visualização.

- Organização de nossa abordagem na forma de um *framework*, incluindo as métricas estruturais usadas e, se possível, um módulo de anotação semântica. Isso facilitaria a geração e teste de estruturas FCA baseadas em nossa abordagem para outros *corpora* e domínios.

- Estudo quanto à inclusão das restrições semânticas associadas aos papéis, nos conceitos formais. Tal inclusão poderia qualificar, ainda mais, os conceitos formais.

- Estudo de outras heurísticas, para reduzir ainda mais a complexidade de construção das estruturas FCA, sem perda de informação semântica relevante.

- Estudo da aplicação de nossa abordagem na realização de outras tarefas, tais como agrupamento de textos, enriquecimento de textos para categorização, suporte a sistemas de pergunta-e-resposta, entre outros.

- Estudo da identificação de instâncias a partir de papéis semânticos e seu uso em estruturas do tipo FCA.

10.5 Produção Científica

Durante o doutoramento, além das publicações em preparação, tivemos a seguinte produção:

- S. Moraes e V. Lima. Um estudo sobre categorização hierárquica de uma grande coleção de textos em língua portuguesa, V Workshop em Tecnologia da Informação e Linguagem Humana (TIL'07), Rio de Janeiro, Brasil, SBC, ed., 2007. (Qualis: -).

- S. Moraes e V. Lima. Abordagem não supervisionada para extração de conceitos a partir de textos, VI Workshop em Tecnologia da Informação e Linguagem Humana (TIL'08), WebMedia, Vila Velha, Espirito Santo, Brasil, SBC, ed., 2008. (Qualis: B3).

- S. Azeredo, S. Moraes e V. Lima. Keywords, k-NN and neural networks: a support for hierarchical categorization of texts in brazilian portuguese, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, ELRA, ed., 2008. (Qualis: A2).

- S. Moraes e V. Lima. Combining Formal Concept Analysis and semantic information for building ontological structures from texts: an exploratory study, Proceedings of International Language Resources and Evaluation (LREC'12), Istanbul, ELRA, ed., 2012. (Qualis: A2).

REFERÊNCIAS BIBLIOGRÁFICAS

[1] E. Agirre, O. Ansa, E. Hovy e D. Martínez. "Enriching Very Large Ontologies Using the WWW". In: 1^{st} Ontology Learning Workshop (OL 2000) at the 14^{th} European Conference on Artificial Intelligence (ECAI'00), Berlin, Germany, 2000, 6p.

[2] H. Alani e C. Brewster. "Metrics for Ranking Ontologies". In: 4^{th} International Workshop on Evaluation of Ontologies for the Web (EON2006) at the 15^{th} International World Wide Web Conference (WWW 2006), Edinburgh, Scotland, 2006, 7p.

[3] E. Alfonseca e S. Manandhar. "An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery". In: 1^{st} International Conference on General Wordnet, Mysore, India, 2002, 9p.

[4] F. Alqadash e R. Bhatnagar. "Similarity Measures in Formal Concept Analysis". *Annals of Mathematics and Artificial Intelligence*, vol. 61-3, Kluwer Academic Publishers, Hingham, MA, USA, Mar 2011, pp. 245–256.

[5] J. An e Y. P. Chen. "Concept Learning of Text Documents". In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), IEEE Computer Society, Washington, DC, USA, 2004, 4p.

[6] M. Antonie e O. R. Zaïane. "Text Document Categorization by Term Association". In: IEEE International Conference on Data Mining (ICDM'02), IEEE Computer Society, Washington, DC, USA, 2002, 8p.

[7] N. Aussenac-Gilles, B. Biebow e S. Szulman. "Revisiting Ontology Design: A Methodology Based on Corpus Analysis". In: 12^{th} European Workshop on Knowledge Acquisition, Modeling and Management (EKAW'00), Springer-Verlag, Juan-les-Pins, France, 2000, 17p.

[8] S. Azeredo, S. Moraes e V. Lima. "Keywords, k-NN and Neural Networks: a Support for Hierarchical Categorization of Texts in Brazilian Portuguese". In: 6^{th} International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008, 6p.

[9] C. F. Baker, C. J. Fillmore e J. B. Lowe. "The Berkeley Framenet Project". In: 17^{th} International Conference on Computational Linguistics, Association for Computational Linguistics, Montreal, Quebec, Canada, 1998, 5p.

[10] C. F. Baker e J. Ruppenhofer. "FrameNet's Frames vs. Levin's Verb Classes". In: 28^{th} Annual Meeting of the Berkeley Linguistics Society, Berkeley, 2002, 12p.

[11] M. Balakrishna, D. Moldovan, M. Tatu e M. Olteanu. "Semi-Automatic Domain Ontology Creation from Text Resources". In: 7^{th} Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010, 8p.

[12] S. L. Bang, J. D. Yang e H. J. Yang. "Hierarchical Document Categorization with k-NN and Concept-based Thesauri". *Information Processing and Management*, vol. 42-2, Pergamon Press, Tarrytown, NY, USA, Mar 2006, pp. 387–406.

[13] T. L. Baségio e V. L. S. Lima. "Semi-automatically Building Ontological Structures from Portuguese Written Texts". In: 7^{th} International Workshop Computational Processing of Portuguese Language (PROPOR 2006), Springer-Verlag, Itatiaia, Brazil, 2006, 4p.

[14] R. Basili, D. Croce, D. Cao e C. Giannone. "Learning Semantic Roles for Ontology Patterns". In: IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'09), IEEE Computer Society, Washington, DC, USA, 2009, 4p.

[15] R. Bendaoud, M. R. Hacene, Y. Toussaint, B. Delecroix e A. Napoli. "Text-based Ontology Construction using Relational Concept Analysis". In: International Workshop on Ontology Dynamics (IWOD 2007), Innsbruck, Austria, 2007, 14p.

[16] R. Bendaoud, A. Napoli e Y. Toussaint. "A Proposal for an Interactive Ontology Design Process based on Formal Concept Analysis". In: 5^{th} International Conference on Formal Ontology in Information Systems (FOIS 2008), IOS Press, Amsterdam, The Netherlands, 2008, 13p.

[17] R. Bendaoud, A. Napoli e Y. Toussaint. "Formal Concept Analysis: A Unified Framework for Building and Refining Ontologies". In: Knowledge Engineering: Practice and Patterns, 16^{th} International Conference on Knowledge Engineering and Knowledge Management (ERAW 2008), Springer, Acitrezza, Italy, 2008, 6p.

[18] K. Bertet, S. Guillas e J. Ogier. "Extensions of Bordat's Algorithm for Attributes". In: 5^{th} International Conference on Concept Lattices and Their Applications (CLA 2007), CEUR-WS.org, Montpellier, France, 2007, 12p.

[19] E. Bick. "Automatic Semantic-Role Annotation for Portuguese". In: V Workshop em Tecnologia da Informação e Linguagem Humana (TIL'07), Anais do XXVII Congresso da SBC, Rio de Janeiro, Brasil, 2007, 4p.

[20] S. Bird. "NLTK: the Natural Language Toolkit". In: COLING/ACL on Interactive Presentation Sessions (COLING-ACL'06), Association for Computational Linguistics, Sydney, Australia, 2006, 4p.

[21] G. Bisson, C. Nédellec e D. Cañamero. "Designing Clustering Methods for Ontology Building - The Mo'K Workbench". In: 1^{st} Workshop on Ontology Learning (OL'00) at the 14^{th} European Conference on Artificial Intelligence (ECAI'00), CEUR-WS.org, Berlin, Germany, 2000, 7p.

[22] E. Blanco e D. Moldovan. "A Model for Composing Semantic Relations". In: 9^{th} International Conference on Computational Semantics (IWCS'11), Association for Computational Linguistics, Oxford, United Kingdom, 2011, 10p.

[23] S. Bloehdorn, P. Cimiano e A. Hotho. "Learning Ontologies to Improve Text Clustering and Classification". In: From Data and Information Analysis to Knowledge Engineering, 29^{th} Annual Conference of the German Classification Society, Springer, Magdeburg, Germany, 2006, 8p.

[24] F. S. Borba. *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil.* Editora da UNESP, São Paulo, 1990, 600p.

[25] J. Brank, M. Grobelnik e D. Mladenić. "Automatic Evaluation of Ontologies". In: Kao, A. e Poteet, S. R. (Ed.), Natural Language Processing and Text Mining, Springer, 2007, 16p.

[26] K.K. Breitmann, M.A. Casanova e W. Truszkowski. *Semantic Web: Concepts, Technologies and Applications*. Springer-Verlag, 2007, 330p.

[27] C. Brewster, H. Alani, S. Dasmahapatra e Y. Wilks. "Data Driven Ontology Evaluation". In: International Conference on Language Resources and Evaluation (LREC'04), European Language Resources Association (ELRA), Lisboa, Portugal, 2004, 5p.

[28] A. Budanitsky e G. Hirst. "Evaluating WordNet-based Measures of Lexical Semantic Relatedness". *Computational Linguistics*, vol. 32-1, MIT Press, Cambridge, MA, USA, Mar 2006, pp. 13–47.

[29] P. Buitelaar, P. Cimiano e B. Magnini. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, 2005, 171p.

[30] P. Buitelaar, D. Olejnik e M. Sintek. "A Protége Plug-in for Ontology Extraction from Text based on Linguistic Analysis", In: 1^{st} European Semantic Web Symposium (ESWS'04), Springer, Heraklion, Crete, Greece, 2004, 14p.

[31] J. Butters e F. Ciravegna. "Using Similarity Metrics for Terminology Recognition". In: 6^{th} International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008, 6p.

[32] W. B. Cavnar e J. M. Trenkle. "N-Gram-Based Text Categorization". In: 3^{rd} Annual Symposium on Document Analysis and Information Retrieval (SDAIR 94), Las Vegas, US, 1994, 15p.

[33] G. Chierchia e S. McConnell-Ginet. *Meaning and grammar: an introduction to semantics*. MIT Press, Cambridge, MA, USA, 2000, 573p.

[34] O. S. Chin, N. Kulathuramaiyer e A. W. Yeo. "Automatic Discovery of Concepts from Text". In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), IEEE Computer Society, Washington, DC, USA, 2006, 4p.

[35] W. C. Cho e D. Richards. "Ontology Construction and Concept Reuse with Formal Concept Analysis for Improved Web Document Retrieval". *Web Intelligence and Agent Systems*, vol. 5-1, IOS Press, Amsterdam, The Netherlands, Jan 2007, pp. 109–126.

[36] I. C. Chow e J. J. Webster. "Integration of Linguistic Resources for Verb Classification: FrameNet Frame, WordNet Verb and Suggested Upper Marged Ontology". In: 8^{th} International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'07), Springer-Verlag, Mexico City, Mexico, 2007, 12p.

[37] J. M. Cigarrán, A. Peñas, J. Gonzalo e F. Verdejo. "Automatic Selection of Noun Phrases as Document Descriptors in an FCA-based Information Retrieval System". In: 3^{th} International Conference on Formal Concept Analysis (ICFCA'05), Springer, Lens, France, 2005, 15p.

[38] P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, Secaucus, NJ, USA, 2006, 347p.

[39] P. Cimiano, A. Hotho e S. Staab. "Clustering Concept Hierarchies from Text". In: 4^{th} International Conference on Language Resources and Evaluation (LREC'04), European Language Resources Association (ELRA), Lisboa, Portugal, 2004, 4p.

[40] P. Cimiano, A. Hotho e S. Staab. "Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text". In: 16^{th} European Conference of Artificial Intelligence (ECAI'04), IOS Press, Valencia, Spain, 2004, 6p.

[41] P. Cimiano, A. Hotho e S. Staab. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research (JAIR)*, vol. 24-1, AAAI Press, USA, Jul 2005, pp. 305-339.

[42] P. Cimiano, S. Staab e J. Tane. "Automatic Acquisition of Taxonomies from Text: FCA meets NLP". In: ECML/PKDD International Workshop on Adaptative Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, 2003, 8p.

[43] P. Cimiano e J. Völker. "Text2Onto - a Framework for Ontology Learning and Data-driven Change Discovery". In: 10^{th} International Conference on Applications of Natural Language to Information Systems (NLDB'05), Springer, Alicante, Spain, 2005, 12p.

[44] C. Comparot, O. Haemmerlé e N. Hernandez. "Conceptual Graphs and Ontologies for Information Retrieval". In: Conceptual Structures: Knowledge Architectures for Smart Applications, 15^{th} International Conference on Conceptual Structures (ICCS'07), Springer, Sheffield, UK, 2007, 4p.

[45] M. Croitoru, B. Hu, S. Dashmapatra, P. H. Lewis, D. Dupplaw e L. Xiao. "A Conceptual Graph based Approach to Ontology Similarity Measure". In: Conceptual Structures: Knowledge Architectures for Smart Applications, 15^{th} International Conference on Conceptual Structures (ICCS'07), Springer, Sheffield, UK, 2007, 11p.

[46] G. Cui, Q. Lu, W. Li e Y. Chen. "Mining Concepts from Wikipedia for Ontology Construction". In: IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology, IEEE, Milan, Italy, 2009, 4p.

[47] H. T. Dang e M. Palmer. "The Role of Semantic Roles in Disambiguating Verb Senses". In: 43^{rd} Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, 8p.

[48] B. A. Davey e H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 1990, 310p.

[49] K. Dellschaft e S. Staab. "On How to Perform a Gold Standard based Evaluation of Ontology Learning". In: International Semantic Web Conference (ISWC'06), Springer, Athens, GA, USA, 2006, 14p.

[50] M. Dittenbach, H. Berger e D. Merkl. "Improving Domain Ontologies by Mining Semantics from Text". In: 1^{st} Asia-Pacific Conference on Conceptual Modelling (APPCCM'04), Conferences in Research and Practice in Information Technology, Dunedin, New Zealand, 2004, 10p.

[51] D. Dowty. Thematic Proto-Roles and Argument Selection. *Language*, vol. 67-3, Linguistic Society of America, Set 1991, pp. 547–619.

[52] A. N. Edmonds. "Using Concept Structures for Efficient Document Comparison and Location". In: IEEE Symposium on Computational Intelligence and Data Mining (CIDM'07), IEEE, Honolulu, Hawaii, USA, 2007, 5p.

[53] M. Ehrig. *Ontology Alignment: Bridging the Semantic Gap (Semantic Web And Beyond Computing for Human Experience)*. Springer-Verlag, 2007, 247p.

[54] P. Eklund e R. Wille. "Semantology as Basis for Conceptual Knowledge Processing". In: 5^{th} International Conference on Formal Concept Analysis (ICFCA'07), Springer-Verlag, Clermont-Ferrand, France, 2007, 21p.

[55] J. Euzenat e P. Shvaiko. *Ontology Matching*. Springer-Verlag, Berlin-Heidelberg, 2007, 333p.

[56] T. J. Everts, S. S. Park e B. H. Kang. "Using Formal Concept Analysis with an Incremental Knowledge Acquisition System for Web Document Management". In: 29^{th} Australasian Computer Science Conference (ACSC'06), Australian Computer Society, Hobart, Australia, 2006, 10p.

[57] I. Falk, C. Gardent e A. Lorenzo. "Using Formal Concept Analysis to Acquire Knowledge about Verbs". In: Concept Lattices and theirs Applications, 7^{th} International Conference on Concept Lattices and Their Applications (CLA'10), Sevilla, Spain, 2010, 12p.

[58] D. Faure e T. Poibeau. "First Experiences of using Semantic Knowledge Learned by ASIUM for Information Extraction Task using INTEX". In: 1^{st} Workshop on Ontology Learning at the 14^{th} European Conference on Artificial Intelligence (ECAI'00), CEUR-WS.org, Berlin, Germany, 2000, 6p.

[59] C. Fellbaum. A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, vol. 32-2, Kluwser Academic Publishers, Mar 1998, pp. 209–220.

[60] E. R. Fernandes, R. L. Milidiú e C. N. Santos. "Portuguese Language Processing Service". In: 18^{th} International World Wide Web Conference Committee (IW3C2'09), ACM, Madrid, Spain, 2009, 7p.

[61] C. J. Fillmore. "The Case for Case". In: E. Bach and R. T. Harms (Ed.), Universals in Linguistic Theory, New York, 1998, 88p.

[62] A. Formica. Ontology-based Concept Similarity in Formal Concept Analysis. *Information Sciences*, vol. 176-18, Elsevier Science Inc., New York, NY, USA, Set 2006, pp. 2624–2641.

[63] A. Formica. Concept Similarity in Formal Concept Analysis: An Information Content Approach. *Knowledge-Based System*, vol. 21-1, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, Fev 2008, pp. 80–87.

[64] B. Fortuna, M. Grobelnik e D. Mladenić. "Semi-automatic Data-driven Ontology Construction System". In: 9^{th} International Multi-conference Information Society (IS'06), Ljubljana, Slovenia, 2006, 4p.

[65] B. Fortuna, D. Mladenić e M. Grobelnik. "Semi-automatic Construction of Topic Ontologies". In: M. Ackermann et. al (Ed), Semantics, Web and Mining, Joint International Workshops, EWMF'05 and KDO'05, Springer, Porto, Portugal, 2006, 11p.

[66] K. T. Frantzi e S. Ananiadou. The C-Value/NC-Value Domain Independent Method for Multi-word Term Extraction. *Journal of Natural Language Processing*, vol. 6-3, 1999, pp. 145-179.

[67] L. C. Freeman. A Set of Measures of Centrality based on Betweenness. *Sociometry*, vol. 40-1, American Sociological Association, Mar 1977, pp. 35–41.

[68] L. A. Freitas e R. Vieira. "Revisão Sistemática sobre Métricas para Ontologias". In: 3^{th} Seminário De Pesquisa Em Ontologia No Brasil (ONTOBRAS'10), Florianópolis, 2010, 12p.

[69] G. Fu e A. Cohn. "Utility Ontology Development with Formal Concept Analysis". In: 5^{th} International Conference on Formal Ontology in Information Systems (FOIS'08), IOS Press, Amsterdam, The Netherlands , 2008, 14p.

[70] H. Fu e E. M. Nguifo. "Lattice Algorithms for Data Mining". In: Revue Ingénierie des Systemes d'Information (ISI'04), Numéro Spécial Extraction et Usages Multiples de Motifs dans les Bases de Données, 2004, 20p.

[71] P. Funk, A. Lewien, G. Snelting, T. Braunschweig e A. Softwaretechnologie. "Algorithms for Concept Lattice Decomposition and their Application", Technical Report, Computer Science Departament, Technische University Braunschweig, 1995, 17p.

[72] P. Gamallo, G. P. Lopes e A. Agustini. "Inducing Classes of Terms from Text". In: 10^{th} International Conference Text, Speech and Dialogue (TSD'07), Springer, Pilsen, Czech Republic, 2007, 8p.

[73] A. Gangemi, C. Catenacci, M. Ciaramita e J. Lehmann. "A Theoretical Framework for Ontology Evaluation and Validation". In: Semantic Web Applications and Perspectives (SWAP'05) , 2^{nd} Italian Semantic Web Workshop, University of Trento, CEUR-WS.org, Trento, Italy, 2005, 16 p.

[74] A. Gangemi, C. Catenacci, M. Ciaramita e J. Lehmann. "Ontology Evaluation and Validation: an Integrated Formal Model for the Quality Diagnostic Task". Technical Report, Laboratory of Applied Ontologies – ISTC-CNR, Rome, Italy, 2005, 53p.

[75] B. Ganter. "Formal Concept Analysis: Algorithmic Aspects ". Technical Report, TU Dresden, Germany, 2002, 53p.

[76] B. Ganter e R. Wille. "Applied Lattice Theory: Formal Concept Analysis". In: General Lattice Theory, G. Grätzer (Ed.), 1997, 14p.

[77] D. Gildea e D. Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics*, vol. 28-3, MIT Press, Cambridge, MA, USA, Set 2002, pp. 245–288.

[78] A. M. Giuglea e A. Moschitti. "Semantic Role labelling via FrameNet, VerbNet and PropBank". In: 21^{st} International Conference on Computational Linguistic and 44^{th} Annual Meeting of the ACL, Association for Computational Linguistics, Sidney , Australia, 2006, 8p.

[79] A. Gómez-Pérez e D. Manzano-Macho. "A Survey of Ontology Learning Methods and Techniques". Technical Report, OntoWeb Consortium, Departamento de Inteligencia Artificial Universidad Politecnica de Madrid, Madrid, Spain, 2003, 86p.

[80] A. Gómez-Pérez e D. Manzano-Macho. An Overview of Methods and Tools for Ontology Learning from Texts. *Knowledge Engineering Review*, vol. 19-3, Cambridge University Press, New York, NY, USA, Set 2004, pp. 187–212.

[81] M. Gonzalez, V. L. S. Lima e J .V. Lima. "Tools for Nominalization: an Alternative for Lexical Normalization". In: 7^{th} International Conference on Computational Processing of the Portuguese Language (PROPOR'06), Springer-Verlag, Itatiaia, Brazil, 2006, 10p.

[82] A. S. Gordon e R. Swanson. "Generalizing Semantic Role Annotations Across Syntactically Similar Verbs". In: 45^{th} Annual Meeting of the Association for Computational Linguistics (ACL'07), The Association for Computer Linguistics, Prague, Czech Republic, 2007, 8p.

[83] G. Grefenstette. "Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window based Approaches". In: Corpus Processing for Lexical Acquisition, Boguraev, B. and Pustejovsky, J. (Ed.), MIT Press, Cambridge, MA, USA, 1996, 12p.

[84] T. Gruber. "Ontology (Computer Science) - Definition in Encyclopedia of Database Systems". In: Encyclopedia of Database System, L. Liu and T. M. Ö (Ed.), Springer-Verlag, 2008. Disponível em: <http://tomgruber.org/writing/ontology-definition-2007.htm>. Acesso em: Jan 2010.

[85] T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, vol. 5-2, Academic Press Ltd., London, UK, UK, Jun 1993, pp. 199–221.

[86] N. Guarino. "Formal Ontology and Information Systems". In: 1^{st} International Conference on Formal Ontologies in Information Systems (FOIS'98) , IOS Press, Trento, Italy, 1998, 13p.

[87] G. Guizzardi. "Ontological foundations for structural conceptual models". Tese de Doutorado, Centre for Telematics and Information Technology, University of Twente, Enschede, The Netherlands, 2005, 416p.

[88] G. Guizzardi, R.A. Falbo e R.S.S. Guizzardi. "A Importância de Ontologias de Fundamentação para a Engenharia de Ontologias de Domínio: o Caso do Domínio de Processo de Software". In: IEEE Latin America Transactions, vol. 6-3, IEEE, Jul 2008, pp. 244-251.

[89] M. R. Hacene, A. Napoli e P. Valtchev. "Ontology Learning from Text using Relational Concept Analysis". In: International MCETECH Conference on e-Technologies (MCETECH'08), IEEE Computer Society, Washington, DC, USA, 2008, 10p.

[90] M. A. K. Halliday e R. Hasan. *Cohesion in English (English Language)*. Longman Pub Group, 1976, 374p.

[91] M. A. Hearst. "Automatic Acquisition of Hyponyms from Large Text Corpora". In: 14^{th} International Conference on Computational Linguistics (COLING'92), Association for Computational Linguistics, Nantes, France, 1992, 7p.

[92] D. Hindle. "Noun Classification from Predicate-Argument Structures". In: 28^{th} Annual Meeting of the Association for Computational Linguistics, Pittsburgh, Pennsylvania, 1990, 8p.

[93] A. Hotho, A. Nürnberger e G. Paass. A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, vol. 20-1, Mai 2005, pp. 19-62.

[94] M. R. Hovav e B. Levin. *Unaccusativity at the Syntax-Lexical Semantics Interface*. MIT Press, Cambridge, MA, 1995, 336p.

[95] X. Hu, X. Wei, D. Wang e P. Li. "A Parallel Algorithm to Construct Concept Lattice". In: 4^{th} International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'07), IEEE Computer Society, Washington, DC, USA, 2007, 5p.

[96] R. Jackendoff. *Semantic Structures*. MIT Press, Cambridge, MA, USA, 1990, 336p.

[97] R. Jackendoff. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, 2002, 477p.

[98] H. Jia, J. Newman e H. Tianfield. "A New Formal Concept Analysis based Learning Approach to Ontology Building". In: 2^{nd} International Conference on Metadata and Semantics Research (MTSR'07), Springer, Corfu Island, Greece, 2007, 12p.

[99] D. Jurafsky e J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* 2^{nd} edition, Prentice Hall, 2009, 988p.

[100] D. Jurkevicius e O. Vasilecas. "Formal Concept Analysis for Concepts Collecting and their Analysis". In: Computer Science and Information Technologies, Latvian University, Riga, 2009, 18p.

[101] V. Kamphuis e J. Sarbo. "Natural Language Concept Analysis". In: Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning (NeMLaP3/CoNLL'98), ACL, Sidney, Australia, 1998, 10p.

[102] S. Kannan e R. Bhaskaran. Association Rule Pruning based on Interestingness Measures with Clustering. *International Journal of Computer Science Issues (IJCSI)*, vol. 6-1, 2009, pp. 35–43.

[103] L. Karoui, M. A. Aufaure e N. Bennacer. "Context-based Hierarchical Clustering for Ontology Learning". In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), IEEE Computer Society, Washington, DC, USA, 2006, 8p.

[104] M. Kavalec e V. Svátek. "A Study on Automated Relation Labelling in Ontology Learning". In: P. Buitelaar, P. Cimmiano e B. Magnini (Ed.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005, 15p.

[105] L. Khan e F. Luo. "Ontology Construction for Information Selection". In: 14^{th} International Conference on Tools with Artificial Intelligence (ICTAI'02), IEEE Computer Society, Washington, DC, USA, 2002, 6p.

[106] J. Kietz, A. Maedche e R. Volz. "A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet", In: 12^{th} European Knowledge Acquisition Workshop (EKAW'00), Springer-Verlag, Juan-les-Pins, France, 2000, 5p.

[107] P. Kingsbury e K. Kipper. "Deriving Verb-Meaning Clusters from Syntactic Structure".In: Workshop on Text Meaning, HLT-NAACL 2003, Association for Computational Linguistics, Morristown, NJ, USA, 2003, 8p.

[108] P. Kingsbury e M. Palmer. "From Treebank to Propbank". In: 3^{rd} International Conference on Language Resources and Evaluation (LREC'02), European Language Resources Association (ELRA), Las Palmas, Gran Canaria, Spain, 2002, 5p.

[109] P. Kingsbury, M. Palmer e M. Marcus. "Adding Semantic Annotation to the Penn TreeBank". In: Human Language Technology Conference (HTL'02), San Diego, California, 2002, 5p.

[110] K. Kipper. "VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon". Tese de Doutorado,University of Pennsylvania, Philadelphia, PA, USA, 2005, 146p.

[111] D. Klein e C. D. Manning. "Fast Exact Inference with a Factored Model for Natural Language Parsing". In: Advances in Neural Information Processing Systems (NIPS'02), MIT Press,Vancouver, British Columbia, Canada, 2003, 8p.

[112] M. Klimushkin, S. A. Obiedkov e C. Roth. "Approaches to the Selection of Relevant Concepts in the Case of Noisy Data". In: 8^{th} International Conference on Formal Concept Analysis (ICFCA'10), Springer-Verlag, Agadir, Morocco, 2010, 12p.

[113] A. Kobayashi, S. Masuyama e S. Sekine. A Method for Automatic Ontology Construction Using Wikipedia. *IEICE TRANSACTIONS on Information and Systems*, vol. J93D-12, 2010, pp. 2597–2609.

[114] A. Korhonen. "Assigning Verbs to Semantic Classes via WordNet". In: Workshop on Building and using Semantic Networks (SEMANET'02), Association for Computational Linguistics, Morristown, NJ, USA, 2002, 7p.

[115] L. Kovics e P. Baranyi. "Document Clustering based on Concept Lattice". In: IEEE International Conference on System, Man and Cybernetics (SMC'02), Hammamet, Tunisia, 2002, 6p.

[116] H. Kuramoto. Sintagmas Nominais: uma Nova Proposta para a Recuperação de Informação. *DataGramaZero - Revista de Ciência da Informação*, vol. 3-1, IASI - Instituto de Adaptação e Inserção na Sociedade da Informação, Rio de Janeiro, Brasil, Fev 2002, 11p.

[117] S. Kuznetsov e S. Obiedkov. Comparing Performance of Algorithms for Generating Concept Lattices. *Journal of Experimental and Theoretical Artificial Intelligence*, vol.14-2-3, 2002, pp. 189–216.

[118] G. Lame e S. Desprès. "Updating Ontologies in the Legal Domain". In: 10^{th} International Conference on Artificial Intelligence and Law (ICAIL'05), ACM, Bologna, Italy, 2005, 8p.

[119] L. C. Langie. Um Estudo sobre a Aplicação do Algoritmo k-NN à Categorização Hierárquica de Textos. Dissertação de Mestrado. Faculdade de Informática, PUCRS, 2004, 126 p.

[120] R. Larson e G. Segal. *Knowledge of Meaning: An Introduction to Semantic Theory.* 2. ed, MIT Press/Bradford Books, Cambridge, MA, USA, 1995, 639p.

[121] R. Y.K. Lau, Y. Li e Y. Xu. "Mining Fuzzy Domain Ontology from Textual Databases". In: IEEE/WIC/ACM International Conference on Web Intelligence (WI '07), IEEE Computer Society, Silicon Valley, USA, 2007, pp.156–162.

[122] C. Leacock e M. Chodorow. *"Combining Local Context and WordNet Similarity for Word Sense Identification".* In: WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Massachusetts, 1998, 18p.

[123] L. Lee. "Measures of Distributional Similarity". In: 37^{th} Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99), Association for Computational Linguistics, College Park, Maryland, 1999, 8p.

[124] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, vol.10-8, Fev 1996, pp. 707–710.

[125] B. Levin. Review of English Verb Classes and Alternations: a Preliminary Investigation. *Computational Linguistics*, vol. 20-3, MIT Press, Cambridge, MA, USA, 1994, pp. 495–497.

[126] B. Levin e M. R. Hovav. *"Lexical Semantics and Syntactic Structure".* In: The Handbook of Contemporary Semantic Theory, Blackwell, Oxford, 1996, 20p.

[127] Y. Li, Y. Yuan, X. Guo, Y. Sheng e L. Chen. "A Fast Algorithm for Generating Concepts". In: International Conference on Information and Automation (ICIA'08), IEEE, Hunan, China, 2008, 6p.

[128] D. Lin. "An Information-Theoretic Definition of Similarity". In: 15^{th} International Conference on Machine Learning (ICML'98), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, 9p.

[129] E. Loper, S. Yi e M. Palmer. "Combining Lexical Resources: Mapping between PropBank and VerbNet". In: 7^{th} International Workshop on Computational Linguistics, Tilburg, The Netherlands, 2007, 12p.

[130] A. Maedche e S. Staab. "Measuring Similarity between Ontologies". In: 13^{th} International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (EKAW '02), Springer-Verlag, Siguenza, Spain, 2002, 13p.

[131] A. Maedche e S. Staab. "Mining Non-Taxonomic Conceptual Relations from Text". In: 12^{th} European Knowledge Acquisition Workshop (EKAW'00), Springer-Verlag, Juan-les-Pins, France, 2000, 12p.

[132] A. Maedche e S. Staab. "Ontology Learning". In: Handbook on Ontologies, Springer, 2004, 26p.

[133] A. Maedche e S. Staab. "The Text-to-Onto Ontology Learning Environment". In: 8^{th} Software Demonstration atInternational Conference on Conceptual Structures (ICCS'00), Darmstadt, Germany, 2000, 5p.

[134] A. Maedche e S. Staab. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, vol. 16-2, IEEE Educational Activities Department, Piscataway, NJ, USA, Mar 2001, pp. 72–79.

[135] C. D. Manning e H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, USA, 1999, 680p.

[136] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz e B. Schasberger. "The Penn Treebank: Annotating Predicate Argument Structure". In: Workshop on Human Language Technology (HLT'94), Association for Computational Linguistics, Plainsboro, NJ, 1994, 6p.

[137] M. P. Marcus, M. A. Marcinkiewicz e B. Santorini. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, vol. 19-2, MIT Press, Cambridge, MA, USA, Jun 1993, pp. 313–330.

[138] M. Marneffe, B. MacCartney e C. D. Manning. "Generating Typed Dependency Parses from Phrase Structure Trees". In: 5^{th} International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006, 6p.

[139] L. Màrquez, X. Carreras, K. C. Litkowski e S. Stevenson. Semantic Role Labeling: an Introduction to the Special Issue. *Computational Linguistics*, vol. 34-2, MIT Press, Cambridge, MA, USA, Jun 2008, pp. 145–159.

[140] N. Meddouri e M. Meddouri. "Classification Methods based on Formal Concept Analysis". In: Concept Lattices and Their Applications, (CLA'08), Palacky University, Olomouc, 2008, 8p.

[141] D. Merwe, S. A. Obiedkov e D. G. Kourie. "AddIntent: A New Incremental Algorithm for Constructing Concept Lattices", In: 2^{nd} International Conference on Formal Concept Analysis (ICFCA'04), Springer, Sydney, Australia, 2004, 14p.

[142] A. Mihis. The Evaluation of Ontology Matching versus Text. *Informatica Economica*, vol. 14-4, 2010, pp. 147–155.

[143] G. A. Miller. WordNet: a Lexical Database for English. *Communications of the ACM*, vol. 38-11, 1995, pp. 39-41.

[144] D. Moldovan e A. Badulescu. "A Semantic Scattering Model for the Automatic Interpretation of Genitives". In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05), Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, 8p.

[145] S. M. W. Moraes e V. L. S. Lima. "Abordagem não supervisionada para Extração de Conceitos a partir de Textos". In: VI Workshop em Tecnologia da Informação e Linguagem Humana (TIL'08), XIV Brazilian Symposium on Multimedia and the Web (WebMedia'08), ACM, Vila Velha, Espirito Santo, Brasil, 2008, 5p.

[146] S. M. W. Moraes e V. L. S. Lima. "Um Estudo sobre Categorização Hierárquica de uma Grande Coleção de Textos em Língua Portuguesa". In: V Workshop em Tecnologia da Informação e Linguagem Humana (TIL'07), SBC, Rio de Janeiro, Brasil, 2007, 10p.

[147] P. Moreda, B. Navarro e M. Palomar. Corpus-based Semantic Role Approach in Information Retrieval. *Data & Knowledge Engineering*, vol. 61-3, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, Jun 2007, pp. 467–483.

[148] A. Moriki e S. Yoshida. "Order-based Clustering using Formal Concept Analysis". In: World Automation Congress (WAC'10), IEEE, Kobe, Japan, 2010, 6p.

[149] N. N. Myat e K. H. S. Hla. "A Combined Approach of Formal Concept Analysis and Text Mining for Concept Based Document Clustering". In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), IEEE Computer Society, Washington, DC, USA, 2005, 4p.

[150] P. Nakov e M. A. Hearst. "Using Verbs to Characterize Noun-Noun Relations". In: 12^{th} International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'06), Springer, Varna, Bulgaria, 2006, 12p.

[151] R. Navigli, P. Velardi e A. Gangemi. Ontology Learning and Its Application to Automated Terminology Translation. *IEEE Intelligent Systems*, vol. 18-1, IEEE Educational Activities Department, Piscataway, NJ, USA, 2003, pp. 22–31.

[152] Y. D. Netzer, D. Gabay, M. Adler, Y. Goldberg e M. Elhadad. "Ontology Evaluation through Text Classification". In: Advances in Web and Network Technologies, and Information Management (APWeb/WAIM'09), Springer-Verlag, Suzhou, China, 2009, 12p.

[153] J. F. Nilsson. "Ontological Constitutions for Classes and Properties". In: Conceptual Structures: Inspiration and Application at 14^{th} International Conference on Conceptual Structures (ICCS'06), Aalborg, Denmark, 2006, 16p.

[154] L. Obrst, W. Ceusters, I. Mani, S. Ray e B. Smith. "The Evaluation of Ontologies: Toward Improved Semantic Interoperability". In: Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, C. J. O. Baker and K. Cheung (Ed.), Springer Verlag. 2006, 20p.

[155] L. J. Old. "Homograph Disambiguation Using Formal Concept Analysis", In: 4^{th} International Conference on Formal Concept Analysis (ICFCA'06), Springer, Dresden, Germany, 2006, 12p.

[156] N. Omar e S. S. Hasbullah. "SRL TOOL: Heuristics-based Semantic Role Labeling through Natural Language Processing". In: IEEE International Symposium on Information Technology 2008 (ITSim'08), IEEE, Kuala Lumpur, Malaysia, 2008, 7p.

[157] P. G. Otero, G. P. Lopes e A. Agustini. Automatic Acquisition of Formal Concepts from Text. *GLDV-Journal for Computational Linguistics and Language Tech nology, LDV Forum - Foundations of Ontologies in Text Technology, Part II : Applications*, vol. 23-1, German Society for Computational Linguistics & Language Technology,Berlin, Germany, 2008, pp. 59–74.

[158] D. E. Pal'chunov. "Lattices of Relatively Axiomatizable Classes". In: 5^{th} International Conference on Formal Concept Analysis (ICFCA'07), Springer, Clermont-Ferrand, France, 2007, 19p.

[159] M. Palmer, D. Gildea e P. Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, vol.31-1, MIT Press, Cambridge, MA, USA, 2005, pp. 71–106.

[160] A. G. Pérez. Ontological Engineering: A State Of The Art. *Expert Update: Knowledge Based Systems and Applied Artificial Intelligence*, vol. 3-2, British Computer Society, 1999, pp. 33-43.

[161] M. A. Perini. *Princípios de Lingüística Descritiva: introdução ao pensamento gramatical.* Parábola Editorial, São Paulo, 2006, 208p.

[162] M. A. Perini. *Gramática Descritiva do Português.* Editora Ática, São Paulo, 2006, 384p.

[163] D. Poshyvanyk e A. Marcus. "Combining Formal Concept Analysis with Information Retrieval for Concept Location in Source Code". In: 15^{th} IEEE International Conference in Program Comprehension (ICPC'07), Banff, Alberta, BC, 2007, 12p.

[164] S. S. Pradhan, W. Ward e J. H. Martin. Towards Robust Semantic Role Labeling. *Computational Linguistics*, vol. 34-2, MIT Press, Cambridge, MA, USA, 2008, pp. 289–310.

[165] S. Prediger. "Logical Scaling in Formal Concept Analysis". In: 5^{th} International Conference on Conceptual Structures: Fulfilling Peirce's Dream (ICCS'97), Springer-Verlag, London, UK, 1997, 10p.

[166] S. Prediger e R. Wille. "The Lattice of Concept Graphs of a Relationally Scaled Context". In: 7^{th} International Conference on Conceptual Structures: Standards and Practices (ICCS'99), Springer-Verlag, London, UK, 1999, 14p.

[167] U. Priss. "Linguistic Applications of Formal Concept Analysis". In: 1^{st} International Conference on Formal Concept Analysis (ICFCA'03), Springer, 2005, 12p.

[168] U. Priss. "Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases". Tese de Doutorado, Technischen Universität Darmstadt, Shaker Verlag, Aachen, 1998, 115p.

[169] U. Priss. The Formalization of WordNet by methods of Relational Concept Analysis. *WordNet: An Eletronic Lexical DataBase and Some of its Applications*, C. Fellbaum (Ed.),MIT Press, 1998, pp. 179–196.

[170] U. Priss. Facet-like Structures in Computer Science. *Axiomathes*, vol. 18-2, Springer Netherlands, 2008, pp. 243-255.

[171] U. Priss. Formal Concept Analysis in Information Science. *Annual Review of Information Science and Technology*, vol.40-1, John Wiley & Sons, Inc., New York, NY, USA, Dez 2006, pp. 521–543.

[172] U. Priss e L. J. Old. "Concept Neighbourhoods in Lexical Databases". In: 8^{th} International Conference on Formal Concept Analysis (ICFCA'10), Springer-Verlag,Agadir, Morocco, 2010, 13p.

[173] U. Priss e L. J. Old. "Data Weeding Techniques Applied to Roget's Thesaurus. Knowledge Processing in Practice". In: 1^{st} International Conference on Knowledge Processing and Data Analysis (KONT'07/KPP'07), Springer-Verlag, Novosibirsk, Russia, 2011, 14p.

[174] V. Punyakanok, D. Roth e W. Yih. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, vol. 34-2, MIT Press, Cambridge, MA, USA, Jun 2008, pp. 257–287.

[175] A. E. Qadi, D. Aboutajedine e Y. Ennouary. Formal Concept Analysis for Information Retrieval. *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 7-2, LJS Publisher and IJCSIS Press, United States, Mar 2010, pp. 119-125.

[176] M. A. Qadir, M. Fahad e M. W. Noshairwan. "On Conceptualization Mismatches Between Ontologies". In: IEEE International Conference on Granular Computing (GRC'07), IEEE Computer Society, Washington, DC, USA, 2007, 4p.

[177] C. Qi, S. Cui e Y. Sun. "Learning Classification Rules Based on Concept Semilattice". In: ISECS International Colloquium on Computing, Communication, Control and Management (CCCM'09), IEEE, Beijing, China, 2009, 5p.

[178] S. Reese, G. Boleda, M. Cuadros, L. Padró e G. Rigau. "Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus". In: 7^{th} Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010, 4p.

[179] P. Resnik. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", In: 14^{th} International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, 6p.

[180] L. C. Ribeiro-Junior. "OntoLP:Construção Semi-automática de Ontologias a partir de Textos da Língua Portuguesa". Dissertação de Mestrado, Programa Interdisciplinar de Pós-Graduação em Computação Aplicada, Universidade do Rio dos Sinos (UNISINOS), São Leopoldo,RS, Brasil, 2008, 158p.

[181] C. Roth, S. A. Obiedkov e D. G. Kourie. On Succinct Representation of Knowledge Community Taxonomies with Formal Concept Analysis. *International Journal Foundations of Computer Science (IJFCS)*, vol. 19-2, World Scientific Publishing, Abr 2008, pp. 383–404.

[182] S. Rudolph, J. Völker e P. Hitzler. "Supporting Lexical Ontology Learning by Relational Exploration". In: Conceptual Structures: Knowledge Architectures for Smart Applications, 15^{th} International Conference on Conceptual Structures (ICCS'07), Springer, Sheffield, UK, 2007, 4p.

[183] R. Haber S. Afonso, E. Bick e D. Santos. "Floresta Sintá(c)tica: a treebank for portuguese". In: 3^{rd} International Conference on Language Resources and Evaluation (LREC'02), European Language Resources Association (ELRA), Las Palmas, Gran Canaria, Spain, 2002, 6p.

[184] P. Saint-Dizier. "An Introduction to the Lexical Semantics of Predicative Forms". In: Predicative Forms in Natural Language and in Lexical Knowledge Bases, Kluwer Academic Publishers, Cambridge, MA, USA, 1999, 52p.

[185] G. Salton e M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, London, U.K., 1983, 448p.

[186] D. Sánchez e A. Moreno. "Discovering Non-taxonomic Relations from the Web". In: 7^{th} International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'06), Springer, Burgos, Spain, 2006, 18p.

[187] D. Sánchez e A. Moreno. Learning Non-taxonomic Relationships from Web Documents for Domain Ontology Construction. *Data & Knowledge Engineering*, vol. 64-3, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, Mar 2008, pp. 600–623.

[188] C. N. Santos. "Entropy Guided Transformation Learning". Tese de Doutorado, Programa de Pós-Graduação do Departamento de Informática, PUC-Rio, Rio de Janeiro, 2009, 85p.

[189] F. Sebastiani. Automatic Classification of Text. *The Encyclopedia of Language and Linguistics*, vol. 14-2, Elsevier Science Publishers, Amsterdam, NL, 2006, pp. 457–462.

[190] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computational Survey*, vol. 34-1, ACM, New York, NY, USA, Mar 2002, pp. 1–47.

[191] M. H. Seddiqui e M. Aono. "Metric of Intrinsic Information Content for Measuring Semantic Similarity in an Ontology". In: 7^{th} Asia-Pacific Conference on Conceptual Modelling (APCCM'10), Australian Computer Society Inc., Brisbane, Australia, 2010, pp. 89–96.

[192] I. Serra e R. Girardi. "Extracting Non-taxonomic Relationships of Ontologies from Texts". In: 6^{th} International Conference Soft Computing Models in Industrial and Environmental Applications (SOCO'11), Springer Berlin, Heidelberg, 2011, 20p.

[193] M. Shamsfard. Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts. *International Journal on Computer Science and Engineering (IJCSE)*, vol. 2-6, Engg Journals Publications, Set 2010, pp. 2190–2196.

[194] S. Shehata, F. Karray e M. Kamel. "A Concept-based Model for Enhancing Text Categorization". In: 13^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07), ACM, San Jose, California, USA, 2007, 9p.

[195] S. Shehata, F. Karray e M. Kamel. "Enhancing Text Clustering Using Concept-based Mining Model". In: 6^{th} International Conference on Data Mining (ICDM'06), IEEE Computer Society, Washington, DC, USA, 2006, 6p.

[196] B. C. D. Silva. "A Construção da Base da Wordnet.Br: Conquistas e Desafios". In: III Workshop em Tecnologia da Informação e Linguagem Humana (TIL'05), XXV Congresso da Sociedade Brasileira de Computação, UNISINOS, São Leopoldo, RS, 2005, 10p.

[197] B. Smith. "Ontology". In: Blackwell Guide to the Philosophy of Computing and Information, L. Floridi (Ed.), Blackwell, 2003, 12p.

[198] B. Smith e C. Welty. "FOIS Introduction: Ontology—Towards a New Synthesis". In: International Conference on Formal Ontology in Information Systems (FOIS'01), ACM Press, Ogunquit, Maine, USA, 2001, 7p.

[199] J. F. Sowa. "Processes and Participants". In: 4^{th} International Conference on Conceptual Structures (ICCS'96), Springer-Verlag, London, UK, 1996, 22p.

[200] I. Spasić, G. Nenadić e S. Ananiadou. "Using Domain-Specific Verbs for Term Classification". In: Workshop on Natural Language Processing in Biomedicine (ACL'03), Association for Computational Linguistics, Sapporo, Japan, 2003, 8p.

[201] G. Stumme. "Exploration Tools in Formal Concept Analysis". In: Studies in Classification, Data Analysis, and Knowledge Organization, Ordinal and Symbolic Data Analysis (OSDA'95), Springer, 1995, 14p.

[202] R. S. Swier e S. Stevenson. "Unsupervised Semantic Role Labelling". In: Conference on Empirical Methods in Natural Language Processing (EMNLP'04), Barcelona, Spain, 2004, 8p.

[203] S. Tamagawa, S. Sakurai, T. Tejima, T. Morita, N. Izumi e T. Yamaguchi. "Learning a Large Scale of Ontology from Japanese Wikipedia". In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'10), IEEE Computer Society, Washington, DC, USA, 2010, 8p.

[204] S. Tartir, B. Arpinar, M. Moore, A. P. Sheth e B. Aleman-Meza. "OntoQA: Metric-Based Ontology Quality Analysis". In: IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, IEEE Computer Society, Houston, Texas, 2005, 9p.

[205] E. Teike e P. Fankhauser. Exploring Lexical Patterns in Text:Lexical Cohesion Analysis with WordNet. *Interdisciplinary Studies on Information Structure (ISIS) - Heterogeneity in Focus: Creating and Using Linguistic Databases*, vol.2-, Berlin, 2005, pp. 129–145.

[206] K. Toutanova, A. Haghighi e C. Manning. A Global Joint Model for Semantic Role Labeling. *Computational Linguistics*, vol. 34-2, MIT Press, Cambridge, MA, USA, Jun 2008, pp. 161–191.

[207] A. D. Troy, G. Zhang e Y. Tian. "Faster Concept Analysis". In: Conceptual Structures: Knowledge Architectures for Smart Applications, 15^{th} International Conference on Conceptual Structures (ICCS'07), Springer, Sheffield, UK, 2007, 14p.

[208] P. Valtchev, M. R. Hacene, M. Huchard e C. Roume. "Extracting Formal Concepts out of Relational Data". In: 4^{th} International Conference Journées de l'Informatique Messine (JIM '03): Knowledge Discovery and Discrete Mathematics, INRIA, Metz, France, 2003, 13p.

[209] P. Valtchev e R. Missaoui. "Similarity-based Clustering versus Galois Lattice Building: Strengths and Weaknesses". In: 14^{th} European Conference on Object-Oriented Programming (ECOOP'00), Sophia Antipolis, Cannes, France, 2000, 10p.

[210] F. J. Valverde-Albacete. Extracting Frame-Semantics Knowledge using Lattice Theory. *Journal of Logic and Computation*, vol. 18-3, Oxford University Press, Oxford, UK, Jun 2008, pp. 361–384.

[211] F. J. Valverde-Albacete e C. Peláez-Moreno. "Galois Connections Between Semimodules and Applications in Data Mining". In: 5^{th} International Conference on Formal Concept Analysis (ICFCA'07), Springer, Clermont-Ferrand, France, 2007, 16p.

[212] P. Velardi, P. Fabriani e M. Missikoff. "Using Text Processing Techniques to Automatically Enrich a Domain Ontology", In: International Conference on Formal Ontology in Information Systems (FOIS'01), ACM, Ogunquit, Maine, USA, 2001, 15p.

[213] J. Völker, P. Haase e P. Hitzler. "Learning Expressive Ontologies". In: Ontology Learning and Population: Bridging the Gap between Text and Knowledge of Frontiers in Artificial Intelligence and Applications, vol. 167, IOS Press, Amsterdam, 2008, 15p.

[214] S. S. Walde. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, vol. 32-2, MIT Press, Cambridge, MA, USA, 2006, pp. 159–194.

[215] A. Weichselbraun, G. Wohlgenannt e A. Scharl. Refining Non-Taxonomic Relation Labels with External Structured Data to Support Ontology Learning. *Data and Knowledge Engineering*, vol. 69-8, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, Ago 2010, pp. 763–778.

[216] A. Weichselbraun, G. Wohlgenannt, A. Scharl, M. Granitzer, T. Neidhart e A. Juffinger. Discovery and Evaluation of Non-Taxonomic Relations in Domain Ontologies. *International Journal of Metadata, Semantics and Ontologies*, vol. 4-3, Inderscience Publishers, Geneva, Switzerland, Ago 2009, pp. 212–222.

[217] C. A. Welty e J. W. Murdock. "Towards Knowledge Acquisition from Information Extraction". In: 5^{th} International Semantic Web Conference (ISWC'06), Springer, Athens, GA, USA, 2006, 14p.

[218] R. Wille. "Conceptual Graphs and Formal Concept Analysis". In: 5^{th} International Conference on Conceptual Structures: Fulfilling Peirce's Dream (ICCS'97), Springer-Verlag, London, UK, 1997, 14p.

[219] R. Wille. "Conceptual Graphs and Formal Concept Analysis". In: 5^{th} International Conference on Conceptual Structures (ICCS'97), Springer, Seattle, Washington, USA, 1997, 14p.

[220] R. Wille. "Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies". In: Formal Concept Analysis, Ganter, B. and Stumme, G. and Wille, R. (Ed.), Springer Berlin, Heidelberg, 2005, 24p.

[221] K. E. Wolff. "A First Course in Formal Concept Analysis – How to Understand Line Diagrams". In: Advances ha Statistical Software, SoftStat'93, F. Faulbaum (Ed.), Gustav Fischer Verlag, Stuttgart, 1994, 10p.

[222] W. Wong, W. Liu e M. Bennamoun. Ontology Learning from Text: A Look back and into the Future. *ACM Computing Surveys*, vol. 44-4, ACM, 2012, 36p.

[223] S. Wu, T. Tsai e W. Hsu. "Text Categorization using Automatically Acquired Domain Ontology". In: 6^{th} International Workshop on Information Retrieval with Asian Languages (AsianIR'03), Association for Computational Linguistics, Sapporo, Japan, 2003, 8p.

[224] Z. Wu e M. Palmer. "Verbs Semantics and Lexical Selection". In: 32^{nd} Annual Meeting on Association for Computational Linguistics (ACL'94), Association for Computational Linguistics, Las Cruces, New Mexico, 1994, 6p.

[225] H. Yang e J. Callan. "Ontology Generation for Large Email Collections". In: Digital Government Research Center, Digital Government Society of North America, Montreal, Canada, 2009, 8p.

[226] Y. Yang e X. Liu. "A Re-examination of Text Categorization Methods". In: 22^{nd} International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), ACM, Berkeley, California, United States, 1999, 8p.

[227] P. Ye e T. Baldwin. "Verb Sense Disambiguation Using Selectional Preferences Extracted with a State-of-the-art Semantic Role Labeler". In: Australasian Language Tecnology Workshop (ALTW'06), Sydney, Australia, 2006, 10p.

[228] J. Yu, J. A. Thom e A. Tam. "Ontology Evaluation using Wikipedia Categories for Browsing". In: 16^{th} ACM Conference on Information and Knowledge Management (CIKM'07), ACM, Lisbon, Portugal, 2007, 10p.

[229] G. Yule. *The Study of Language*. Cambridge University Press, Cambridge, 1996, 294p.

[230] E. Zavitsanos, G. Paliouras e G.A. Vouros. "Discovering Subsumption Hierarchies of Ontology Concepts from Text Corpora". In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), IEEE Computer Society, Washington, DC, USA, 2007, 7p.

[231] G. Zhang, A. D. Troy e K. Bourgoin. "Bootstrapping Ontology Learning for Information Retrieval Using Formal Concept Analysis and Information Anchors". In: 14^{th} International Conference on Conceptual Structures, Aalborg, Denmark, 2006, 14p.

[232] Y. Zhang e B. Feng. Clustering Search Results Based on Formal Concept Analysis. *Information Technology Journal*, vol. 7-5, Asian Network for Scientific Information, Pakistan, 2008, pp. 746–753.

[233] Z. Zhang, J. Iria, C. Brewster e F. Ciravegna. "A Comparative Evaluation of Term Recognition Algorithms". In: 6^{th} International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008, 3p.

[234] W. Zhou, Z. Liu e Y. Zhao. "Concept Hierarchies Generation for Classification using Fuzzy Formal Concept Analysis". In: 8^{th} International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'07), IEEE Computer Society, Washington, DC, USA, 2007, 6p.

APÊNDICE A - Conjuntos ordenados e reticulados

Davey e Priestley em [48] descrevem ordenação como uma relação binária sobre um conjunto de objetos. Formalmente, uma *ordem* (ou *ordem parcial*) sobre um conjunto P é uma relação binária $\leq$ em P, tal que para todo $x, y, z \in P$, as seguintes propriedades são satisfeitas:

- reflexividade: $x \leq x$;
- anti-simetria: $x \leq y$ e $y \leq x$ implica $x = y$;
- transitividade: $x \leq y$ e $y \leq z$ implica $x \leq z$.

Um conjunto P que possui uma relação de ordem $\leq$ definida para seus elementos, denotada por $(P; \leq)$, é dito conjunto ordenado. Essa ordenação pode ser total ou parcial. Em uma ordenação total, também conhecida por cadeia *(chain)*, quaisquer dois elementos do conjunto são comparáveis, ou seja, para todo $x, y \in P$, ou $x \leq y$ ou $y \leq x$. Um exemplo de conjunto totalmente ordenado é $(\mathbb{N}; \leq)$ onde a relação "é menor ou igual a" existe para quaisquer dois naturais: $0 \leq 1 \leq 2 \leq 3 \leq ...$ Já a relação "é divisor de" para $\mathbb{N}$ representada por nós como $\preceq$, estabelece uma ordem parcial $(\mathbb{N}; \preceq)$, pois $m \preceq n$ se e somente se existe um $k \in \mathbb{N}$, tal que $km = n$.

Os conjuntos ordenados podem ser representados por diagramas de Hasse. Nesses diagramas os elementos de um conjunto ordenado são representados por círculos e a relação de ordem, por linhas interconectadas. Na Figura A.1 há cinco exemplos de conjuntos ordenados. O diagrama

- A.1(a) representa um subconjunto dos $\mathbb{N}$, a cadeia $(\{0, 1, 2, 3, 4\}, \leq)$;
- A.1(b) descreve o conjunto dos divisores de 24 $(\{1, 2, 3, 4, 6, 12, 24\}, \preceq)$;
- A.1(c) corresponde ao conjunto $(\wp(\{a, b, c\}), \subseteq)$, onde $\wp$ é o conjunto potência[1] de $\{a, b, c\}$;
- A.1(d) representa um subconjunto dos termos usados na área jurídica($\{civil, penal, direito, direito\_civil, direito\_penal\}$,$é\_substring\_de$); e
- A.1(e) descreve um subconjunto dos tipos de obras de arte[2] existentes ($\{poema, crônica, pintura, fotografia, música, teatro, obra\_de\_arte, obra\_literária, obra\_performática, obra\_visual\}$, $é\_subclasse\_de$).

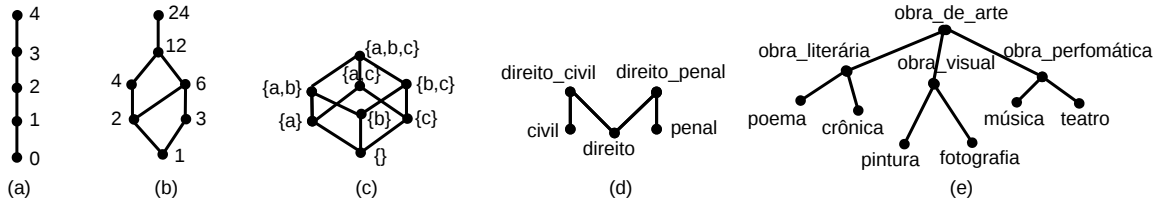


Figura A.1 – Exemplos de conjuntos ordenados

[1] O conjunto potência de um conjunto A é o conjunto das partes de A, ou seja, todos os subconjuntos finitos que podem ser formados a partir de A.

[2] Esse diagrama é um subconjunto da ontologia descrita em http://www.inf.pucrs.br/~ontolp/Visualizacao/Arte/28102.html

Um conjunto ordenado possui propriedades que são definidas a partir da existência de limites superiores e inferiores de seus subconjuntos. Sendo P um conjunto ordenado e S um subconjunto de P ($S \subseteq P$), um elemento $x \in P$ é um limite (cota) superior de S se $s \leq x$ para todo $s \in S$. Por exemplo, para $\{2, 3\}$, que é subconjunto de A.1(b), os limites superiores são os elementos $\{6, 12, 24\}$. Para o subconjunto $\{\{\}, \{a\}, \{c\}, \{a, c\}\}$ de A.1(c), os limites superiores correspondem a $\{\{a, c\}, \{a, b, c\}\}$. Em A.1($e$), os limites superiores do subconjunto $\{fotografia, pintura\}$ são $\{obra\_visual, obra\_de\_arte\}$. Já no caso de $\{direito\_civil, direito\_penal\}$, que é subconjunto de A.1 (d), não há limites superiores.

Um limite inferior é justamente o oposto, ocorre quando $s \geq x$ para todo $s \in S$. Para o subconjunto $\{2, 3\}$ de A.1(b), $\{1\}$ é o limite inferior; para $\{\{\}, \{a\}, \{c\}, \{a, c\}\}$ de A.1(c) o limite inferior é $\{\{\}\}$; e para $\{direito\_civil, direito\_penal\}$ de A.1 (d), $\{direito\}$ é o limite inferior. No entanto, no caso de $\{fotografia, pintura\}$ de A.1(e) não há limites inferiores.

Um conjunto ordenado é chamado de reticulado (*lattice*) quando pode-se definir para todos os seus subconjuntos elementos supremo e ínfimo. Considerando a relação de ordem do conjunto, o supremo é o menor elemento das suas cotas superiores e o ínfimo, o maior elemento das suas cotas inferiores. Enquanto as cotas podem ser conjuntos, o supremo e ínfimo, quando existem, são únicos. Para o subconjunto $\{4, 6\}$ do diagrama A.1(b), as cotas superiores são $\{12, 24\}$ e as inferiores $\{1, 2\}$, portanto o seu supremo é 12 e seu ínfimo, 2. Se considerarmos todo conjunto A.1(b), o supremo é 12 e o ínfimo, 1.

Quando os conjuntos ordenados não possuem supremo ou infímo, são chamados de semirreticulados. Se a estrutura possuir apenas o supremo denomina-se semirreticulado superior e se possuir apenas o ínfimo, semirreticulado inferior.

Dos conjuntos ordenados apresentados na Figura A.1 apenas A.1(d) não é reticulado ou semirreticulado, pois não possui nem supremo e nem ínfimo. O diagrama A.1(e) é um semirreticulado superior e todos os demais são reticulados.

Outros conceitos importantes são maximais e minimais. Um elemento x é um maximal se $\nexists y$ tal que $y > x$. Um elemento minimal é o oposto, x é um minimal se $\nexists y$ tal que $y < x$. No diagrama A.1(d), os elementos $\{direito\_civil\}$ e $\{direito\_penal\}$ são maximais, assim como $\{civil\}$, $\{direito\}$ e $\{penal\}$ são minimais. No caso de um reticulado, há apenas um elemento maximal e um minimal, os quais costumam ser chamados, respectivamente, de elemento *top* ($\top$) e *bottom* ($\bot$). No diagrama A.1(c), o elemento *top* é $\{a, b, c\}$ e o *bottom*, $\{\}$.

APÊNDICE B - Similaridade máxima de atributos para medida Sim

Neste apêndice são mostrados os cálculos realizados para determinar os valores da medida Sim nos exemplos que são apresentados na Seção 3.7.2.

No primeiro exemplo, o objetivo é calcular a $Sim(C_1, C_2)$, onde $C_1 = (\{share, stock, index\}, \{buy, sell, rise\})$ e $C_2 = (\{share, stock, fund\}, \{buy, decline, invest\})$. Para isso, precisamos determinar o valor de s, que é parte da medida Sim, definida em (9). Inicialmente formamos os pares de atributos a partir de $\{buy, sell, rise\} \times \{buy, decline, invest\}$, o qual gera o conjunto $A = \{(buy, buy), (buy, decline), (buy, invest), (sell, buy), (sell, decline), (sell, invest), (rise, buy), (rise, decline), (rise, invest)\}$.

Para formar o conjunto P , onde $P \subseteq A$, escolhemos os pares de A que não possuem atributos em comum e a soma de suas medidas ics seja máxima. A Tabela B.1 mostra a aplicação da ics aos pares de A. O cálculo da medida ics foi realizado com o pacote NLTK (Seção 5.3.2), usando-se a WordNet 3.0 como taxonomia e o Brown como *corpus* de referência. Desta forma, os pares escolhidos para P são $\{(buy, buy), (sell, invest), (rise, decline)\}$, pois valor de s é $1,79$, que o máximo de similaridade obtida entre os atributos dos conceitos C_1, e C_2.

Tabela B.1 – Aplicando a ics aos pares do conjunto A formado pelos atributos dos conceitos C_1 e C_2 .

buy	*sell*	*rise*
ics(buy,buy)= 1	*ics(sell,buy) = 0,06*	*ics(rise,buy) = 0,32*
ics(buy,decline)= 0,06	*ics(sell,decline) = 0.,05*	***ics(rise,decline) = 0,79***
ics(buy,invest) = 0	***ics(sell,invest) = 0***	*ics(rise,invest) = 0*

No segundo exemplo, o objetivo é calcular a $Sim(C_1, C_3)$, onde $C_3 = (\{share, stock, market, rate\}, \{rise, decline\})$. Para definir o valor de s, construimos o conjunto B resultante de $\{buy, sell, rise\} \times \{rise, decline\}$, ou seja, $\{(buy, rise), (buy, decline), (sell, rise), (sell, decline), (rise, rise), (rise, decline)\}$. Para formar o conjunto P , escolhemos os pares $\{(buy, decline), (rise, rise)\}$, pois a soma de suas ics é máxima, sendo assim, $s = 1,06$ (Tabela B.2).

Tabela B.2 – Aplicando a ics aos pares do conjunto B formado pelos atributos dos conceitos C_1 e C_3 .

buy	*sell*	*rise*
ics(buy,decline)= 0,06	*ics(sell,decline) = 0.05*	*ics(rise,decline) = 0.79*
ics(buy,rise) = 0,32	*ics(sell,rise) = 0*	***ics(rise,rise) = 1***

APÊNDICE C - Dados complementares do processamento dos corpora Penn TreeBank Sample e SemLink 1.1

Neste apêndice são mostrados dados extraídos durante o pré-processamento e alinhamentos dos *corpora* Penn TreeBank Sample e SemLink 1.1.

A Tabela C.1 exibe a quantidade de termos relacionados a cada etiqueta POS[3]. Dos 100.673 *tokens* encontrados, 6.592 não tinham etiquetas POS válidas, estavam marcados com *-NONE-*.

Tabela C.1 – Quantidade de *tokens* associados a cada etiqueta POS no *corpus* Penn TreeBank Sample

POS	#	**POS**	#	**POS**	#
NN	13.166	VBZ	2.125	JJS	182
IN	9.857	PRP	1.716	WRB	178
NNP	9.409	VBG	1.460	RBR	136
DT	8.165	VBP	1.321	-RRB-	126
-NONE-	6.592	MD	927	-LRB-	120
NNS	6.047	POS	824	EX	88
JJ	5.834	PRP$	766	RBS	35
,	4.886	$	724	PDT	27
.	3.872	"	712	#	16
CD	3.546	’	694	WP$	14
VBD	3.043	:	563	LS	13
RB	2.822	WDT	445	FW	4
VB	2.554	JJR	381	UH	3
CC	2.265	NNPS	244	SYM	1
TO	2.179	WP	241		
VBN	2.134	RP	216		

No caso dos verbo, identificamos 245 classes VerbNet (Figura **C.1**), sendo que 178 correspondem a classes principais e 67, a subclasses.

Contabilizamos as classes conforme as instâncias de seus verbos. A Tabela C.2 apresenta as 123 classes que possuem mais de 10 instâncias.

Na Tabela C.3, apresentamos os verbos que encontramos para as 5 classes VerbNet mais frequentes. Juntamente com os verbos, mostramos a quantidade de instâncias desses verbos que encontramos nos textos.

A Tabela C.4 descreve o significado de 22 papéis temáticos da VerbNet. O significado desses papéis foi definido com base nas informações constantes na página da própria VerbNet (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html#thetaroles). E a Tabela C.5 apresenta os 28 papéis semânticos identificados e suas respectivas frequências. Esses 28 papéis correspondem aos 22 descritos anteriormente e suas variantes.

As informações extraídas dos *corpora* Penn TreeBank Sample e SemLink 1.1 Sample referentes aos verbos e seus argumentos foram estruturadas em tuplas, tal como é exemplificado na Tabela C.6.

[3]O significado de cada etiqueta POS pode ser encontrado em http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Classes principais: 29.3, 37.7, 27, 48.1.2, 29.2, 55.1, 48.1.1, 29.1, 11.3, 37.5, 30.1, 14, 51.7, 13.6, 48.2, 3, 62, 65, 31.2, 37.6, 13.5.1, 37.1, 63, 29.7, 9.3, 9.5, 45.4, 30.2, 35.3, 29.4, 1, 45.6, 64, 10.5, 43.4, 13.3, 13.4.2, 15.1, 90, 31.3, 55.2, 9.4, 21.1, 51.4.1, 26.4, 89, 13.5.2, 47.8, 51.3.2, 33, 67, 11.4, 40.1.2, 46, 29.5, 39.1, 31.1, 54.1, 50, 25.4, 15.2, 54.2, 68, 31.4, 37.1.1, 51.2, 54.4, 34, 26.6.2, 59, 10.1, 9.8, 35.6, 32.2, 51.3.1, 51.6, 36.1, 9.1, 25.1, 22.2, 22.4, 9.10, 13.4.1, 32.1, 54.3, 48.3, 11.1, 10.2, 2, 85, 25.3, 37.2, 37.8, 36.3, 26.1, 10.6, 76, 47.5.2, 61, 35.2, 47.2, 47.7, 87, 35.5, 58, 47.6, 18.3, 78, 37.4, 10.7, 37.10, 13.2, 23.2, 41.3.1, 25.2, 52, 29.8, 12, 40.5, 35.4, 9.2, 42.1, 37.9, 47.4, 51.8, 16, 20, 23.3, 26.6, 81, 79, 37.3, 45.5, 30.3, 42.2, 29.6, 23.1, 9.9, 26.2, 45.1, 10.4.1, 11.2, 71, 51.4.2, 77, 9.6, 40.2, 54.5, 70, 40.4, 40.3.3, 44, 17.1, 18.4, 24, 39.7, 93, 47.3, 53.1, 43.2, 26.5, 40.8.4, 41.1.1, 40.3.2, 38, 45.2, 84, 30.4, 37.1.2, 45.3, 40.8.2, 39.4, 40.7, 35.1, 18.1, 49, 19, 9.7

Subclasses: 22.1-2, 26.4-1, 26.1-1, 51.1-1, 13.1-1, 47.1-1-1, 47.1-1, 22.1-1, 9.2-1, 11.2-1-1, 51.1-2, 45.6-1, 13.5.2-1, 15.1-1, 13.1-2, 31.3-8, 37.1-1, 22.2-3-1, 29.9-1-1-1, 21.1-1, 36.3-1, 22.3-2, 40.8.3-2, 26.7-1, 9.1-2, 9.3-2, 17.1-1-1, 13.5.1-1, 11.1-1, 17.1-1, 55.1-1, 26.3-1, 9.6-1, 9.7-2, 42.1-1, 26.7-1-1, 21.2-2, 9.7-1, 29.8-1, 10.3-1, 22.2-1, 13.2-1, 36.4-1, 37.9-1, 12-1, 29.5-1, 40.3.1-1, 37.11-1-1, 26.6.2-1, 47.8-1, 11.5-1, 31.3-1, 22.3-1-1, 36.1-1, 10.4.1-1, 29.9-1-1, 22.2-2-1, 18.4-1, 39.3-2, 13.4.1-1, 40.1.2-1, 18.1-1, 47.5.1-2, 55.2-1, 21.2-1, 37.1-1-1, 13.4.2-1

Figura C.1 – Classes VerbNet identificadas no SemLink Sample.

Tabela C.2 – As 123 classes VerbNet mais frequentes nos *corpora* analisados.

Classe	#	**Classe**	#	**Classe**	#	**Classe**	#	**Classe**	#
37.7	1.727	9.4	66	11.1-1	37	22.1-1	21	25.4	15
45.4	332	13.6	63	13.2	37	22.2-3-1	21	37.8	15
45.6	278	31.3	62	37.2	36	37.9-1	21	51.3.1	15
13.1-1	258	51.3.2	62	3	35	29.5-1	20	85	15
13.5.1	215	9.1	58	36.1	33	37.6	20	13.2-1	14
55.1	208	47.8	56	9.8	32	87	20	34	14
1	201	10.5	55	21.1	31	35.4	19	43.2	14
29.4	196	13.5.2-1	55	26.6.2	31	37.1	19	15.1-1	13
29.5	196	59	55	15.1	30	47.7	19	35.2	13
13.3	167	10.1	54	2	30	48.2	19	51.1-1	13
29.2	148	13.5.1-1	53	30.2	30	50	19	54.2	13
62	139	35.6	53	55.2	29	15.2	18	63	13
31.1	119	13.4.1	52	90	29	25.1	18	22.4	12
51.1-2	115	22.2	52	25.2	28	26.1	18	26.6	12
32.1	102	33	52	26.7-1	28	37.5	18	37.11-1-1	12
37.1-1	102	67	51	11.3	27	68	18	47.2	12
29.1	100	22.1-2	50	54.5	27	31.3-8	17	54.3	12
47.1-1	91	31.2	50	14	26	32.2	17	10.2	11
48.1.1	91	26.4-1	49	51.7	26	13.4.2	16	11.4	11
65	88	26.4	46	89	26	17.1-1	16	48.1.2	11
26.1-1	86	76	45	9.10	25	36.3	16	52	11
54.4	86	27	43	48.3	24	51.2	16	64	11
30.1	75	61	41	51.6	24	78	16	9.7-1	11
29.3	71	11.1	40	29.6	23	9.2-1	16		
13.5.2	70	54.1	39	37.4	22	23.2	15		

Tabela C.3 – As 5 classes VerbNet mais frequentes no Penn TreeBank Sample

Classe VerbNet	**Verbos**
37.7	*to say (1.533), to propose (27), to announce (24), to suggest (26), to claim (20), to disclose (20), to report (16), to respond (12), to insist (15), to declare (8), to state (8), to observe (5), to mention (4), to reply (3), to insinuate (2), to remark (2), to utter (1), to voice (1).*
45.4	*to close (41), to improve (24), to increase (24), to grow (19), to operate (18), to expand (17), to slow (16), to open (14), to advance (12), to ease (11), to change (10), to mature (9), to revive (6), to weaken (6), to clear (5), to diversity (4), to reopen (4), to stretch(4), to divide (3), to double (3), to heat (3), to shut (3), to sink (3), to sweeten (3), to vary (3), to accelerate (2), to air (2), to broaden (2), to centralize(2), to cool (2), to diminish (2), to fill (2), to flood (2), to freeze (2), to halt (2), to inflate (2), to lengthen (2), to lessen (2), to level (2), to moderate (2), to narrow(2), to publicize (2), to tighteh (2), to triple (2), to abate (1), to atter (1), to blur (1), to brighten (1), to burn (1), to chill (1), to contract (1), to demobilize (1), to dissolve (1), to evapore (1), to fade (1), to fatten (1), to firm (1), to heighten (1), to magnify (1), to mobilize (1), to multiply (1), to mute (1), to polarize (1), to rarefy (1), to rekindle (1), to reverse (1), to ripen (1), to soften (1), to strengthen (1), to taper (1), to unwind (1), to ventilate (1), to worsen (1).*
45.6	*to rise (94), to fall (45), to increase (33), to decline (18), to drop (14), to gain (13), to jump (11), to climb (10), to grow (9), to surge (9), to soar (8), to plunge (7), to skyrocket (2), to tumble (2), to diminish (1), to decrease (1), to plummet (1).*
13.1-1	*to sell (99), to give (75), to pay (72), to refund (5), to repay (4), to lend (2), to pass (1).*
13.5.1	*to buy (85), to win (27), to reach (25), to find (21), to earn (18), to gain (12), to hire (8), to catch (4), to secure (3), to book (2), to call (2), to gather (2), to reserve (1), to fetch (1), to lease (1), to shoot (1), to leave (1), to pick (1).*

Tabela C.4 – 22 papéis temáticos VerbNet encontrados no SemLink 1.1 Sample.

Papel Temático	**Descrição**
Actor	é usado para algumas classes de comunicação, quando ambos os argumentos podem ser considerados simétricos (pseudo-agentes).
Agent	indica geralmente uma pessoa ou algo animado.
Asset	é utilizado para indicar posse de dinheiro (moeda).
Attribute	corresponde ao atributo de um Patient ou Theme, que está sendo alterado.
Beneficiary	corresponde à entidade que se beneficia de alguma ação.
Cause	é usado principalmente por classes envolvendo verbos com aspectos psicológicos e verbos que são relativos ao corpo.
Destination	é usado para localização espacial, corresponde ao ponto final de um movimento ou à direção desse movimento.
Experiencer	é usado para um participante que tem conhecimento ou experimenta alguma coisa.
Extent	é utilizado para especificar um intervalo ou grau de mudança.
Instrument	indica objetos (ou forças) que entram em contato com algum outro objeto e causam mudança neles.
Location	é usado para localização espacial, indica origem, destino ou lugar.
Material	é usado por classes de verbos relativos à criação e à transformação, indica o ponto de início de transformação.
Patient	é utilizado para os participantes que estão passando por um processo ou que tenham sido afetados por ele de alguma forma.
Predicate	é utilizado para as classes com um complemento predicativo.
Product	é usado por classes de verbos relativos à criação e à transformação, indica o resultado final de uma transformação.
Proposition	indica uma proposta ou oferta.
Recipient	indica o alvo de uma transferência.
Source	é usado para localização espacial, indica o ponto de partida de um movimento.
Stimulus	é utilizado por verbos relativos à percepção, indicam eventos ou objetos que provocam alguma resposta a partir de um Experiencer.
Theme	refere-se aos participantes de uma situação ou àqueles que estão passando por uma mudança de posição.
Topic	é usado por verbos de comunicação, indica o tema ou assunto de uma conversa ou mensagem.
Value	indica o valor de algo.

Tabela C.5 – Frequência dos papéis semânticos associados aos termos do Penn TreeBank Sample.

Papel	#	**Papel**	#	**Papel**	#
Actor1	32	Location	61	Stimulus	45
Actor2	19	Material	14	Theme	1.862
Agent	1.560	Patient	576	Theme1	91
Asset	18	Patient1	96	Theme2	73
Attribute	34	Patient2	51	Topic	394
Beneficiary	11	Predicate	167	Value	10
Cause	144	Product	144		
Destination	91	Proposition	4		
Experiencer	179	Recipient	124		
Extent	88	Source	50		

Tabela C.6 – Exemplos de relações entre os verbo e seus argumentos, juntamente com as respectivas informações semânticas

		argumento1			**argumento2**			
verbo	**classe do verbo**	**nome**	**papel semântico Verbnet**	**papel semântico PropBank**	**nome**	**papel semântico Verbnet**	**papel semântico PropBank**	**fonte: n° sentença, texto**
buy	13.5.1	share	Theme	ARG1	shareholder	Agent	ARG0	02, wsj_0073
buy	13.5.1	share	Asset	ARG3	shareholder	Source	ARG2	00, wsj_0151
mail	11.1-1	company	Agent	ARG0	shareholder	Destination	ARG2	00, wsj_0073
offer	13.3	company	Agent	ARG0	share	Recipient	ARG2	03, wsj_0063
pay	13.1-1	company	Agent	ARG0	share	Theme	ARG1	02, wsj_0150
receive	13.5.2	share	Theme	ARG1	stockholder	Agent	ARG0	22, wsj_0018
receive	13.5.2	share	Theme	ARG1	shareholder	Agent	ARG0	12, wsj_0063
say	37.7	dividend	Topic	ARG1	analyst	Agent	ARG0	05, wsj_0090
sell	13.1-1	share	Theme	ARG1	stockholder	Agent	ARG0	26, wsj_0090

APÊNDICE D - Dados complementares à análise I do Estudo II referente à representação de informações semânticas em conceitos formais

Neste apêndice estão os dados usados em estudos preliminares a partir dos quais realizamos a análise I do Estudo II. Este estudo referia-se à representação de informações semânticas em conceitos formais.

Tabela D.1 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 2: sem sementes, corte 3 (4 papéis semânticos)

caso	**relação I (g,m)**	**#objetos**	**#atributos**	**#conceitos**	**SSM_W**	**SSM_L**	**média**
1	(sn,v)	91	43	66	0,22	0,10	0,16
2	(sn,psV)	91	4	9	0,43	0,56	**0,50**
3	(sn,psV_sn)	91	111	87	0,06	0,03	0,05
4	(sn,cV)	91	29	54	0,27	0,19	0,23
5	(sn,psV)+(sn,cV)	91	33	112	0,30	0,24	0,27
6	(sn,psV_sn)+(sn,cV)	91	148	134	0,14	0,12	0,13

Tabela D.2 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 2: sem sementes, corte 3 (4 papéis semânticos).

caso	**relação I (g,m)**	**#arestas**	**#unitários (%)**	**altura**	**largura**
1	(sn,v)	129	19 (28,8)	4	32
2	(sn,psV)	13	0	3	4
3	(sn,psV_sn)	160	62 (71,3)	4	66
4	(sn,cV)	104	16 (29,6)	5	22
5	(sn,psV)+(sn,cV)	247	28 (25,0)	7	36
6	(sn,psV_sn)+(sn,cV)	263	71 (53,0)	5	75

Tabela D.3 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 3: sem sementes, corte 4 (4 papéis semânticos).

caso	**relação I (g,m)**	**#objetos**	**#atributos**	**#conceitos**	**SSM_W**	**SSM_L**	**média**
1	(sn,v)	50	26	42	0,24	0,13	0,19
2	(sn,psV)	50	4	10	0,32	0,35	**0,34**
3	(sn,psV_sn)	50	62	50	0,06	0,04	0,05
4	(sn,cV)	50	19	35	0,28	0,16	0,22
5	(sn,psV)+(sn,cV)	50	23	73	0,33	0,22	0,28
6	(sn,psV_sn)+(sn,cV)	50	81	81	0,16	0,13	0,15

Tabela D.4 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 3: sem sementes, corte 4 (4 papéis semânticos).

caso	**relação I (g,m)**	**#arestas**	**#unitários (%)**	**altura**	**largura**
1	(sn,v)	81	12 (28,6)	4	18
2	(sn,psV)	15	1 (10,0)	4	3
3	(sn,psV_sn)	90	33 (66,0)	3	36
4	(sn,cV)	67	10 (28,6)	5	14
5	(sn,psV)+(sn,cV)	158	14 (19,2)	6	24
6	(sn,psV_sn)+(sn,cV)	161	41 (50,6)	5	43

Tabela D.5 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 5: 10 sementes, corte 2 (4 papéis semânticos).

caso	**relação I (g,m)**	**#objetos**	**#atributos**	**#conceitos**	**SSM_W**	**SSM_L**	**média**
1	(sn,v)	131	57	91	0,17	0,13	0,15
2	(sn,psV)	131	4	12	0,32	0,23	**0,28**
3	(sn,psV_sn)	131	149	57	0,06	0,07	0,07
4	(sn,cV)	131	35	72	0,24	0,16	0,20
5	(sn,psV)+(sn,cV)	131	39	135	0,30	0,23	0,27
6	(sn,psV_sn)+(sn,cV)	131	184	171	0,15	0,19	0,17

Tabela D.6 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 5: 10 sementes, corte 2 (4 papéis semânticos).

caso	**relação I (g,m)**	**#arestas**	**#unitários (%)**	**altura**	**largura**
1	(sn,v)	179	16 (17,6)	5	52
2	(sn,psV)	19	1 (8,3)	4	4
3	(sn,psV_sn)	102	24 (42,1)	4	28
4	(sn,cV)	149	13 (18,1)	6	32
5	(sn,psV)+(sn,cV)	322	17 (12,6)	7	49
6	(sn,psV_sn)+(sn,cV)	388	44 (25,7)	6	64

Tabela D.7 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 1: sem sementes, corte 2 (todos os papéis).

caso	**relação I (g,m)**	**#arestas**	**#unitários (%)**	**altura**	**largura**
1	(sn,v)	644	90 (31,1)	5	130
2	(sn,psV)	192	21 (25,0)	6	23
3	(sn,psV_sn)	678	208 (58,4)	5	233
4	(sn,cV)	630	79 (30,3)	6	104
5	(sn,psV)+(sn,cV)	1.570	106 (18,0)	7	198
6	(sn,psV_sn)+(sn,cV)	1.384	228 (37,7)	6	262

Tabela D.8 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 2: sem sementes, corte 3 (todos os papéis).

caso	relação I (g,m)	#objetos	#atributos	#conceitos	SSM_W	SSM_L	média
1	(sn,v)	216	112	203	0,23	0,11	0,17
2	(sn,psV)	216	19	66	0,25	0,15	0,20
3	(sn,psV_sn)	216	348	229	0,10	0,06	0,08
4	(sn,cV)	216	68	188	0,25	0,13	0,19
5	(sn,psV)+(sn,cV)	216	87	411	0,32	0,21	**0,27**
6	(sn,psV_sn)+(sn,cV)	216	416	398	0,20	0,15	0,18

Tabela D.9 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 2: sem sementes, corte 3 (todos os papéis).

caso	relação I (g,m)	#arestas	#unitários (%)	altura	largura
1	(sn,v)	439	63 (31,0)	5	92
2	(sn,psV)	147	21 (31,8)	6	23
3	(sn,psV_sn)	441	144 (62,9)	5	156
4	(sn,cV)	437	57 (30,3)	6	73
5	(sn,psV)+(sn,cV)	1.067	83 (20,2)	7	139
6	(sn,psV_sn)+(sn,cV)	898	167 (42,0)	6	178

Tabela D.10 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 3: sem sementes, corte 4 (todos os papéis)

caso	relação I (g,m)	#objetos	#atributos	#conceitos	SSM_W	SSM_L	média
1	(sn,v)	143	91	152	0,24	0,13	0,19
2	(sn,psV)	143	18	56	0,27	0,17	0,22
3	(sn,psV_sn)	143	234	160	0,06	0,05	0,06
4	(sn,cV)	143	59	136	0,26	0,14	0,20
5	(sn,psV)+(sn,cV)	143	77	297	0,33	0,23	**0,28**
6	(sn,psV_sn)+(sn,cV)	143	293	274	0,20	0,15	0,18

Tabela D.11 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 3: sem sementes, corte 4 (todos os papéis)

caso	relação I (g,m)	#arestas	#unitários (%)	altura	largura
1	(sn,v)	314	48 (31,6)	5	73
2	(sn,psV)	125	19 (33,9)	5	20
3	(sn,psV_sn)	301	107 (66,9)	4	113
4	(sn,cV)	297	43 (31,6)	5	52
5	(sn,psV)+(sn,cV)	301	66 (22,2)	4	113
6	(sn,psV_sn)+(sn,cV)	596	119 (43,4)	5	125

Tabela D.12 – Resultados da medida SSM para as estruturas FCA geradas a partir da forma de seleção 5: 10 sementes, corte 2 (todos os papéis).

caso	**relação I (g,m)**	**#objetos**	**#atributos**	**#conceitos**	**SSM_W**	**SSM_L**	**média**
1	(sn,v)	266	100	165	0,21	0,13	0,17
2	(sn,psV)	266	18	62	0,33	0,23	0,28
3	(sn,psV_sn)	266	357	152	0,06	0,03	0,05
4	(sn,cV)	266	66	150	0,27	0,20	0,24
5	(sn,psV)+(sn,cV)	266	84	331	0,34	0,26	**0,30**
6	(sn,psV_sn)+(sn,cV)	266	423	350	0,16	0,16	0,16

Tabela D.13 – Dados estruturais complementares das estruturas FCA geradas a partir da forma de seleção 5: 10 sementes, corte 2 (todos os papéis)

caso	**relação I (g,m)**	**#arestas**	**#unitários (%)**	**altura**	**largura**
1	(sn,v)	352	32 (19,4)	5	94
2	(sn,psV)	132	15 (24,2)	10	17
3	(sn,psV_sn)	286	66 (43,4)	4	85
4	(sn,cV)	342	28 (18,7)	6	71
5	(sn,psV)+(sn,cV)	856	39 (11,8)	7	121
6	(sn,psV_sn)+(sn,cV)	797	95 (27,1)	6	139

APÊNDICE E - Dados complementares à análise II do Estudo II referente à representação de informações semânticas em conceitos formais

Neste apêndice estão os resultados completos da análise II relativas ao estudo de heurísticas usadas nos casos de estudo proposto. Esta análise faz parte do Estudo II no qual investigamos a representação de informações semânticas em conceitos formais.

Tabela E.1 – Resultados da medida SSM para caso $1_{(\text{sn, v})}$ após uso de heurísticas de agrupamento e corte: forma de seleção 5 (todos os papéis)

k	m	#rel$_N$	#rel$_E$	#objetos	#atributos	#conceitos	#unitários (%)	SSM$_W$	SSM$_L$	média
4	3	1.225	42	248	79	145	28 (19,3)	0,33	0,16	0,25
4	4	1.225	42	248	79	145	28 (19,3)	0,33	0,16	0,25
4	5	1.225	126	215	58	109	19 (17,4)	0,39	0,18	**0,29**
5	3	1.225	42	248	79	145	28 (19,3)	0,33	0,16	0,25
5	4	1.225	42	248	79	145	28 (19,3)	0,33	0,16	0,25
5	5	1.225	126	215	58	109	19 (17,4)	0,39	0,18	**0,29**
6	3	1.225	42	248	79	145	28 (19,3)	0,33	0,16	0,25
6	4	1.225	42	248	79	145	28 (19,3)	0,33	0,16	0,25
6	5	1.225	126	215	58	109	19 (17,4)	0,39	0,18	**0,29**

Tabela E.2 – Resultados da medida SSM para caso $2_{(\text{sn, psV})}$ após uso de heurísticas de agrupamento e corte: forma de seleção 5 (todos os papéis)

k	m	#rel$_N$	#rel$_E$	#objetos	#atributos	#conceitos	#unitários (%)	SSM$_W$	SSM$_L$	média
4	3	0	6	266	15	59	13 (22,0)	0,32	0,25	**0,29**
4	4	0	9	265	14	55	11 (20,0)	0,32	0,24	0,28
4	5	0	13	263	13	53	11 (20,8)	0,33	0,25	**0,29**
5	3	0	6	266	15	59	13 (22,0)	0,32	0,25	**0,29**
5	4	0	9	265	14	55	11 (20,0)	0,32	0,24	0,28
5	5	0	13	263	13	53	11 (20,8)	0,33	0,25	**0,29**
6	3	0	6	266	15	59	13 (22,0)	0,32	0,25	**0,29**
6	4	0	9	265	14	55	11 (20,0)	0,32	0,24	0,28
6	5	0	13	263	13	53	11 (20,8)	0,33	0,25	**0,29**

Tabela E.3 – Resultados da medida SSM para caso $3_{(\mathrm{sn},\ \mathrm{ps}V_s\mathrm{n})}$ após uso de heurísticas de agrupamento e corte: forma de seleção 5 (todos os papéis)

k	**m**	**#rel$_N$**	**#rel$_E$**	**#objetos**	**#atributos**	**#conceitos**	**#unitários (%)**	**SSM$_W$**	**SSM$_L$**	**média**
4	3	44	359	194	46	60	14 (23,3)	0,08	0,15	0,12
4	4	44	419	154	26	34	8 (23,6)	0,12	0,27	0,20
4	5	44	435	140	22	28	6 (21,4)	0,15	0,28	**0,22**
5	3	60	358	194	47	61	14 (22,9)	0,08	0,14	0,11
5	4	60	418	154	27	36	9 (25,0)	0,13	0,23	0,18
5	5	60	434	140	23	31	8 (25,8)	0,15	0,23	0,19
6	3	77	357	194	48	61	14 (22,9)	0,08	0,14	0,11
6	4	77	417	154	28	36	9 (25,0)	0,14	0,23	0,19
6	5	77	433	140	24	31	8 (25,8)	0,15	0,23	0,19

Tabela E.4 – Resultados da medida SSM para caso $4_{(\mathrm{sn},\ \mathrm{cV})}$ após uso de heurísticas de agrupamento e corte: forma de seleção 5 (todos os papéis)

k	**m**	**#rel$_N$**	**#rel$_E$**	**#objetos**	**#atributos**	**#conceitos**	**#unitários (%)**	**SSM$_W$**	**SSM$_L$**	**média**
4	3	728	22	257	55	138	26 (18,8)	0,36	0,22	0,29
4	4	728	22	257	55	138	26 (18,8)	0,36	0,22	0,29
4	5	728	78	234	41	114	22 (19,3)	0,39	0,26	**0,33**
5	3	728	22	257	55	138	26 (18,8)	0,36	0,22	0,29
5	4	728	22	257	55	138	26 (18,8)	0,36	0,22	0,29
5	5	728	78	234	41	114	22 (19,3)	0,39	0,26	**0,33**
6	3	728	22	257	55	138	26 (18,8)	0,36	0,22	0,29
6	4	728	22	257	55	138	26 (18,8)	0,36	0,22	0,29
6	5	728	78	234	41	114	22 (19,3)	0,39	0,26	**0,33**

APÊNDICE F - Dados complementares ao Estudo III quanto à tarefa de categorização de textos

Neste apêndice estão os resultados completos quanto à investigação em categorização de textos realizada no Estudo III, por meio da qual avaliamos a aplicabilidade da proposta.

Nas tabelas apresentadas, q indica a quantidade de regras extraídas de cada estrutura ou ontologia; TP corresponde à quantidade de verdadeiros-positivos; FP, à quantidade de falsos-positivos; FN, à quantidade de falso-negativos; Pr, à medida precisão; Re , à medida *recall* e $F1$, à medida que combina Pr e Re. As medidas Pr, Re e $F1$ são tradicionalmente usadas para avaliar os resultados em categorização de textos [190].

Tabela F.1 – Resultados da categorização por regras extraídas das estruturas $\text{TourismFCA}_{\text{caso1}}$ e $\text{FinanceFCA}_{\text{caso1}}$ para o conjunto $\text{teste}_{\text{Wiki}}$

	Tourism						**Finance**						**Macro-Médias**		
q	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
23	138	79	20	0,64	0,87	0,74	81	20	79	0,80	0,50	0,62	0,72	0,69	0,68
24	140	80	18	0,64	0,89	0,74	80	18	80	0,82	0,50	0,62	0,73	0,69	0,68
25	141	80	17	0,64	0,89	0,74	80	17	80	0,82	0,50	0,62	**0,73**	**0,70**	**0,68**
26	140	82	18	0,63	0,89	0,74	78	18	82	0,81	0,49	0,61	0,72	0,69	0,67
27	140	80	18	0,64	0,89	0,74	80	18	80	0,82	0,50	0,62	0,73	0,69	0,68

Tabela F.2 – Resultados da categorização por regras extraídas das estruturas $\text{TourismFCA}_{\text{caso1}}$ e $\text{FinanceFCA}_{\text{caso1}}$ para o conjunto $\text{teste}_{\text{Wiki+PTBS}}$

	Tourism						**Finance**						**Macro-Médias**		
q	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
23	138	106	20	0,57	0,87	0,67	93	20	106	0,82	0,47	0,60	0,69	0,67	0,64
24	140	109	18	0,56	0,89	0,69	90	18	109	0,83	0,45	0,59	0,70	0,67	0,64
25	141	103	17	0,58	0,89	0,70	96	17	103	0,85	0,48	0,62	**0,71**	**0,69**	**0,66**
26	140	104	18	0,57	0,89	0,70	95	18	104	0,84	0,48	0,61	0,71	0,68	0,65
27	140	102	18	0,58	0,89	0,70	97	18	102	0,84	0,49	0,62	**0,71**	**0,69**	**0,66**

Tabela F.3 – Resultados da categorização por regras extraídas das estruturas $\text{TourismFCA}_{\text{caso7}}$ e $\text{FinanceFCA}_{\text{caso7}}$ para o conjunto $\text{teste}_{\text{Wiki}}$

	Tourism						**Finance**						**Macro-Médias**		
q	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
23	149	18	9	0,89	0,94	0,92	142	9	18	0,94	0,89	0,91	0,92	0,92	0,92
24	148	17	10	0,90	0,94	0,92	143	10	17	0,93	0,89	0,91	0,92	0,92	0,92
25	148	16	10	0,90	0,94	0,92	144	10	16	0,94	0,90	0,92	**0,92**	**0,92**	**0,92**
26	146	16	12	0,90	0,92	0,91	144	12	16	0,92	0,90	0,91	0,91	0,91	0,91
27	146	17	12	0,90	0,92	0,91	143	12	17	0,92	0,89	0,91	0,91	0,91	0,91

Tabela F.4 – Resultados da categorização por regras extraídas das estruturas TourismFCA$_{\text{caso7}}$ e FinanceFCA$_{\text{caso7}}$ para o conjunto teste$_{\text{Wiki+PTBS}}$

	Tourism						Finance						Macro-Médias		
q	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
23	149	84	9	0,64	0,94	0,76	115	9	84	0,93	0,58	0,71	0,78	0,76	0,74
24	148	75	10	0,66	0,94	0,78	124	10	75	0,93	0,62	0,74	0,79	0,78	0,76
25	148	70	10	0,68	0,94	0,79	129	10	70	0,93	0,65	0,76	**0,80**	**0,79**	**0,78**
26	146	72	12	0,67	0,92	0,78	127	12	72	0,91	0,64	0,75	0,79	0,78	0,77
27	146	75	12	0,66	0,92	0,77	124	12	75	0,91	0,62	0,74	0,79	0,77	0,76

Tabela F.5 – Resultados da categorização por (todas as) regras extraídas das ontologias de Turismo e Finanças para o conjunto teste$_{Wiki}$

	Tourism						Finance								
regras	**TG = 28**			**T = 14**			**F = 41**			**L = 22**			**Macro-médias**		
ontologias	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
O_{TG}+ O_F	60	11	98	0,84	0,38	0,52	149	98	11	0,60	0,93	**0,73**	**0,72**	**0,66**	**0,63**
O_T+ O_F	135	90	23	0,60	0,85	**0,70**	70	23	90	0,75	0,44	0,55	0,68	0,65	**0,63**
O_{TG}+ O_L	28	4	130	0,88	0,18	0,29	156	130	4	0,55	0,98	0,70	0,71	0,98	0,50
O_T+ O_L	65	21	93	0,76	0,41	0,53	139	93	21	0,60	0,87	0,71	0,68	0,64	0,62

Tabela F.6 – Resultados da categorização por (todas as) regras extraídas das ontologias de Turismo e Finanças para o conjunto teste$_{Wiki+PTBS}$

	Tourism						Finance								
regras	**TG = 28**			**T = 14**			**F = 41**			**L = 22**			**Macro-médias**		
ontologias	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
O_{TG}+ O_F	60	0	98	1,0	0,38	0,55	199	98	0	0,67	1,0	0,80	**0,84**	**0,69**	**0,68**
O_T+ O_F	135	197	23	0,41	0,85	0,55	2	23	197	0,08	0,01	0,02	0,24	0,43	0,28
O_{TG}+ O_L	28	0	130	1,0	0,18	0,30	199	130	0	0,60	1,0	0,75	0,80	0,59	0,53
O_T+ O_L	65	10	93	0,87	0,41	0,56	189	93	10	0,67	0,95	0,79	0,77	0,68	0,67

Tabela F.7 – Resultados da categorização por regras extraídas das ontologias TGPROTON e Finance para o conjunto teste$_{Wiki}$

#regras		Tourism						Finance						Macro-médias		
TG	**F**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
6	6	84	38	74	0,69	0,53	0,60	122	74	38	0,62	0,76	0,69	0,66	0,65	0,64
7	7	89	38	69	0,70	0,56	0,62	122	69	38	0,64	0,76	0,70	0,67	0,66	0,66
8	8	89	37	69	0,71	0,56	0,63	123	69	37	0,64	0,77	0,70	0,67	0,67	0,66
9	9	87	28	71	0,76	0,55	0,64	132	71	28	0,65	0,82	0,72	**0,70**	**0,69**	**0,68**
10	10	85	41	73	0,67	0,54	0,60	119	73	41	0,62	0,74	0,67	0,65	0,64	0,64
11	11	77	39	81	0,66	0,49	0,56	121	81	39	0,60	0,76	0,67	0,63	0,62	0,62
12	12	77	34	81	0,69	0,49	0,57	126	81	34	0,60	0,79	0,69	0,65	0,64	0,63
13	13	75	34	83	0,69	0,47	0,56	126	83	34	0,60	0,79	0,68	0,65	0,63	0,62

Tabela F.8 – Resultados da categorização por regras extraídas das ontologias TGPROTON e Finance para o conjunto teste$_{Wiki+PTBS}$

#regras		Tourism						Finance						Macro-médias		
TG	**F**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
6	6	84	7	74	0,92	0,53	0,67	192	74	7	0,72	0,96	0,82	0,82	0,75	0,75
7	7	89	9	69	0,91	0,56	0,69	190	69	9	0,73	0,95	0,83	0,82	0,76	0,76
8	8	89	6	69	0,94	0,56	0,70	193	69	6	0,74	0,97	0,84	**0,84**	**0,77**	**0,77**
9	9	87	5	71	0,95	0,55	0,70	194	71	5	0,73	0,97	0,83	0,84	0,76	0,77
10	10	85	8	73	0,91	0,54	0,68	191	73	8	0,72	0,96	0,83	0,82	0,75	0,75
11	11	77	7	81	0,92	0,49	0,64	192	81	7	0,70	0,96	0,81	0,81	0,73	0,73
12	12	77	4	81	0,95	0,49	0,64	195	91	4	0,71	0,98	0,82	0,83	0,73	0,73
13	13	75	2	83	0,97	0,47	0,64	197	83	2	0,70	0,99	0,82	0,84	0,73	0,73

Tabela F.9 – Resultados da categorização por regras extraídas das ontologias de Turismo (TG+T) e Finanças (F+L) para o conjunto teste$_{Wiki}$

#regras p/	Tourism (TG+T)						Finance (F+L)						Macro-médias		
ontologia	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
6	94	31	64	0,75	0,59	0,66	129	64	31	0,67	0,81	0,73	0,71	0,70	0,70
7	88	26	70	0,77	0,56	0,65	134	70	26	0,66	0,84	0,74	0,71	0,70	0,69
8	94	28	64	0,77	0,59	0,67	132	64	28	0,67	0,82	0,74	0,72	0,71	0,71
9	92	24	66	0,79	0,58	0,67	136	66	24	0,67	0,85	0,75	**0,73**	**0,72**	**0,71**
10	89	33	69	0,72	0,56	0,64	127	69	33	0,65	0,79	0,71	0,69	0,68	0,67
11	83	34	75	0,71	0,52	0,60	126	75	34	0,63	0,79	0,70	0,67	0,66	0,65
12	85	20	73	0,75	0,54	0,63	132	73	28	0,64	0,83	0,72	0,70	0,68	0,68
13	80	25	78	0,76	0,51	0,61	135	78	25	0,63	0,84	0,72	0,70	0,68	0,67

Tabela F.10 – Resultados da categorização por regras extraídas das ontologias de Turismo (TG+T) e Finanças (F+L) para o conjunto teste$_{Wiki+PTBS}$

#regras p/	Tourism (TG+T)						Finance (F+L)						Macro-médias		
ontologia	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
6	94	10	64	0,90	0,59	0,72	189	64	10	0,75	0,95	0,84	**0,83**	**0,77**	**0,78**
7	88	17	70	0,84	0,56	0,67	182	70	17	0,72	0,91	0,81	0,78	0,74	0,74
8	94	28	64	0,77	0,59	0,67	171	64	28	0,73	0,86	0,79	0,75	0,73	0,73
9	92	33	66	0,74	0,58	0,65	166	66	33	0,72	0,83	0,77	0,73	0,71	0,71
10	89	56	69	0,61	0,56	0,59	143	69	56	0,67	0,72	0,70	0,64	0,64	0,64
11	83	69	75	0,55	0,52	0,54	130	75	69	0,63	0,65	0,64	0,59	0,59	0,59
12	85	110	73	0,44	0,54	0,48	89	73	110	0,55	0,45	0,49	0,49	0,49	0,49
13	80	103	78	0,44	0,51	0,47	96	78	103	0,55	0,48	0,51	0,49	0,49	0,49

Tabela F.11 – Resultados da categorização por k-NN do conjunto teste$_{\text{Wiki}}$, usando seleção por *rank* para *n=50*

	Tourism						Finance						Macro-médias		
k	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
7	148	11	10	0,93	0,94	0,93	149	10	11	0,94	0,93	0,93	0,93	0,93	0,93
13	150	11	8	0,93	0,95	0,94	149	8	11	0,95	0,93	0,94	0,94	0,94	0,94
17	151	9	7	0,94	0,96	0,95	151	7	9	0,96	0,94	0,95	**0,95**	**0,95**	**0,95**

Tabela F.12 – Resultados da categorização por k-NN do conjunto teste$_{\text{Wiki+PTBS}}$, usando seleção por *rank* para *n=50*

	Tourism						Finance						Macro-médias		
k	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
7	148	56	10	0,73	0,94	0,81	143	10	56	0,93	0,72	0,81	0,83	0,83	0,82
13	150	53	8	0,74	0,95	0,83	146	8	53	0,95	0,74	0,83	0,84	0,84	0,83
17	151	50	7	0,75	0,96	0,84	149	7	50	0,96	0,75	0,84	**0,85**	**0,85**	**0,84**

Tabela F.13 – Resultados da categorização por k-NN do conjunto teste$_{\text{Wiki}}$, usando seleção por *rank* para *n=100*

	Tourism						Finance						Macro-medias		
k	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
7	146	10	12	0,94	0,93	0,93	150	12	10	0,93	0,94	0,93	0,93	0,93	0,93
13	147	6	11	0,96	0,93	0,95	154	11	6	0,93	0,96	0,95	**0,95**	**0,95**	**0,95**
17	147	8	11	0,95	0,93	0,94	152	11	8	0,93	0,95	0,94	0,94	0,94	0,94

Tabela F.14 – Resultados da categorização por k-NN do conjunto teste$_{\text{Wiki+PTBS}}$, usando seleção por *rank* para *n=100*

	Tourism						Finance						Macro-médias		
k	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
7	146	32	12	0,82	0,92	0,87	167	12	32	0,93	0,84	0,88	0,88	0,88	0,88
13	147	27	11	0,84	0,93	0,89	172	11	27	0,94	0,86	0,90	**0,89**	**0,90**	**0,89**
17	147	28	11	0,84	0,93	0,88	171	11	28	0,94	0,86	0,90	0,89	0,89	0,89

Tabela F.15 – Resultados da categorização por k-NN do conjunto teste$_{\text{Wiki}}$, usando seleção por *rank* para *n=150*

	Tourism						Finance						Macro-médias		
k	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
7	148	5	10	0,97	0,94	0,95	155	10	5	0,94	0,97	0,96	**0,95**	**0,95**	**0,95**
13	148	8	10	0,95	0,94	0,94	152	10	8	0,94	0,95	0,94	0,94	0,94	0,94
17	146	8	12	0,95	0,92	0,94	152	12	8	0,93	0,95	0,94	0,94	0,94	0,94

Tabela F.16 – Resultados da categorização por k-NN do conjunto teste$_{\text{Wiki+PTBS}}$, usando seleção por *rank* para *n=150*

	Tourism						Finance						Macro-médias		
k	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**TP**	**FP**	**FN**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
7	148	25	10	0,86	0,94	0,89	174	10	25	0,95	0,84	0,90	0,90	0,90	0,90
13	148	24	10	0,86	0,94	0,90	175	10	24	0,95	0,88	0,91	0,90	0,91	0,90
17	146	18	12	0,89	0,92	0,91	181	12	18	0,94	0,91	0,92	**0,91**	**0,92**	**0,92**

ANEXO A - Algoritmo para calcular estabilidade dos conceitos

Este anexo apresenta o algoritmo para calcular a estabilidade de conceitos apresentada na Seção 3.6.3. A versão original desse algoritmo foi definida por Roth *et al.* em [181] e aplica-se apenas ao cálculo da estabilidade intensional. Os operadores $>$ e $\succ$ referem-se, respectivamente, à relação de ordem superconceito-subconceito e à relação de vizinho menor. Um conceito (C, D) é um vizinho menor de (A, B), se $(C, D) < (A, B)$ e não existe um conceito (E, F) tal que $(C, D) < (E, F) < (A, B)$.

O cálculo da estabilidade extensional é realizado de forma equivalente.

Algoritmo CalculaEstabilidadeIntensional{
  $Conceitos = \mathscr{B}(G, M, I)$
  Para cada (A, B) em $Conceitos$ {
    $Contador[(A, B)]$= número de vizinhos menores que o conceito(A, B)
    $Subconjuntos[(A, B)]=2^{|A|}$
  }
  while *Conceitos* não é vazio{
    Seja (C, D) qualquer conceito do conjunto *Conceitos* com $Contador[(C, D)] = 0$
    $Estabilidade[(C, D)] = Subconjuntos[(C, D)]/2^{|C|}$
    remove o conceito (C, D)do conjunto $Conceitos$
    Para cada $(A, B) > (C, D)$ {
      $Subconjuntos[(A, B)] = Subconjuntos[(A, B)] - Subconjuntos[(C, D)]$
      se $(A, B) \succ (C, D)$ então{
        $Contador[(A, B)] = Contador[(A, B)] - 1$
      }
    }
  }
}