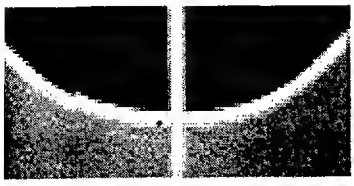

UNIVERSIDADE DE BRASILIA
FACULDADE DE ESTUDOS SOCIAIS APLICADOS
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO E DOCUMENTAÇÃO

AVALIAÇÃO DA SIMILARIDADE SEMÂNTICA ENTRE CLASSES DE ENTIDADES ESPACIAIS, REPRESENTADAS NUMA ONTOLOGIA *AD-HOC*

Por

Paulo César Rodrigues Borges

Tese submetida ao Departamento de Ciência da Informação e Documentação, como requisito parcial para a obtenção do grau de Doutor em Ciência da Informação.

Orientadora

Profª Suzana Pinheiro Machado Mueller

Brasília, julho de 2003.

UNIVERSIDADE DE BRASILIA
FACULDADE DE ESTUDOS SOCIAIS APLICADOS
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO E DOCUMENTAÇÃO
CURSO DE DOUTORADO EM CIÊNCIA DA INFORMAÇÃO

AVALIAÇÃO DA SIMILARIDADE SEMÂNTICA ENTRE CLASSES DE ENTIDADES ESPACIAIS, REPRESENTADAS NUMA ONTOLOGIA *AD-HOC*

Orientadora: SUZANA PINHEIRO MACHADO MUELLER - Profª e Doutora

Coorientadora: MARISA BRÄSCHER BASÍLIO MEDEIROS - Profª e Doutora

Doutorando: PAULO CÉSAR RODRIGUES BORGES

Brasília - DF

2003

Tese apresentada ao Departamento de Ciência da Informação e Documentação da Universidade de Brasília como requisito parcial para obtenção do grau de Doutor.

Brasília, 31 de julho de 2003.

AAprovado por:

PProfª Drª. Suzana Pinheiro Machado Mueller - Presidente

PProfª Drª. Marisa Bräscher Basílio Medeiros - Membro

PProfª Drª. Sely Maria de Souza Costa - Membro

PProf. Dr. Joacil Basílio Rael - Membro

PProf. Dr. Plácido Flaviano Curvo Filho – Membro

PProf. Dr. Antônio Lisboa Carvalho de Miranda - Suplente

In memorian

"A memória olha para o passado. A nova consciência olha para o futuro. O espaço é um dado fundamental nesta descoberta."

(Mílton Santos, geógrafo e professor emérito da USP – 1926 a 2001)

Este trabalho recebeu apoio da 1ª Divisão de Levantamento (RS), do Centro de Cartografia Automatizada do Exército (DF) e da empresa SISGRAPH (SP), representante da INTERGRAPH, Inc., no Brasil.

À Sheyla, Caio, Filipe, Lívia, Úrsula e Tiago,
razões da minha obstinação.

AGRADECIMENTOS

À Profª Suzana, minha orientadora e, ouso dizer, amiga, pela valiosa orientação oferecida, sempre enxergando o que me fugia à vista, tanto em nível genérico como em nível minucioso, sabendo contrariar sem magoar o seu orientando, porque dosa, com maestria, rigor científico e afabilidade. Agradeço-lhe, especialmente, professora Suzana, pela confiança que depositou em mim e pelo apoio emocional e moral que me dispensou nos momentos mais cruciais desta empresa, o que me incentivou e fortaleceu na conquista deste objetivo de vida.

À professora Marisa, minha coorientadora, que secundou a orientação com foco persistente nos objetivos específicos e na orladura das ontologias, repassou-me conselhos que muito tempo me pouparam, mercê de sua tarimbada experiência em PLN, adquirida no seu doutorado, parcialmente realizado na França e consagrada por sua passagem em várias funções do primeiro escalão do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), entre as quais se inclui a sua atual direção.

Ao meu amigo, Dr. Plácido, pelas observações de ordem prática na área de SIG, o que propiciou o norteamento da tese não só pelos aspectos de ordem teórica, mas, nomeadamente, nas aplicações do dia-a-dia da Engenharia Cartográfica.

Aos meus estimados amigos e professores Rael e Sely Maria, que tanto me ajudaram na crítica do texto do projeto, com suas orientações precisas e empréstimo de material didático, altruísmo que às vezes lhes custou caro em termos de tempo livre.

Aos meus colegas e professores componentes de banca, Murilo, Mamede, Moresi Paulo Roberto, Sidnei, Samir e Antônio Miranda, pela gentileza em terem aceito o meu convite.

Ao meu caro professor e amigo, Antônio Miranda, por ter cedido amavelmente o seu escritório de estudos no Departamento de Ciência da Informação, que se tornou uma extensão do meu lar. Este refúgio foi decisivo para levar a bom termo etapas da pesquisa em que a tranqüilidade não era mais uma opção dispensável. Esta saleta pôs-me distante do tumulto e até da realidade do mundo para conceber os requisitos das minhas "criaturas intelectuais em gestação", entre elas, o PRONTO©, ferramenta fundamental para fechar a metodologia.

Aos diversos pesquisadores estrangeiros e nacionais, pela desinteressada cooperação com o meu trabalho. Citá-los seria um exercício difícil, pelo pouco espaço disponível nesta seção, sobre o qual já começo a abusar, sem contar que eu poderia estar cometendo injustiças com as omissões. Apesar disso, não poderia deixar de registrar o que fez por mim a minha colega distante, Maria Andrea Rodríguez Tastets, professora da Universidade de Concepción (Chile), PhD do Curso de Engenharia e Ciência da Informação Espacial da Universidade do Maine (EUA), pelo precioso, ininterrupto (até meados de 2001) e amigável contacto que manteve comigo, especialmente por correio eletrônico, o que me ajudou a dar continuidade ao seu trabalho de pesquisa, aqui, no Brasil.

Ao meu estimado ex-comandante do 10º Regimento de Cavalaria (MS), Gen. Lima, que foi protagonista em dois momentos decisivos da minha vida acadêmica: em 1985, quando eu era um tenente, por me homenagear com uma formatura de todo o regimento pela minha aprovação no concurso de admissão para o IME e, em 1999, eu, já no posto de tenente-coronel e o senhor, como Secretario Interino de Ciência e Tecnologia do Exército, por me conceder a licença em tempo integral de três anos para cursar o doutorado na UnB, em que pese as resistências encontradas.

Ao Sr. Ex-Diretor do Serviço Geográfico do Exército, Gen. Armindo, por ter-me indicado para o curso em apreço.

Ao atual Diretor do Serviço Geográfico do Exército, prezado Gen. Paiva de Sá, pelo apoio prestado em concordar com duas prorrogações ao prazo de conclusão da tese e pelas providências administrativas que mandou tomar para que a aplicação dos questionários aos engenheiros e topógrafos da DSG e do CCAuEx ocorresse em tempo hábil e sem prejuízos ao cronograma já apertado da pesquisa.

Aos meus companheiros da 1ª Divisão de Levantamento e, em especial, ao seu chefe, Ten-Cel Monteiro Soares, por torcer pelo meu êxito e pelo apoio material em transporte e hospedagem que arcou por mim durante o XX Congresso Brasileiro de Cartografia (Porto Alegre, 2001), período que foi muito proveitoso para a elaboração do primeiro protótipo que orientou a metodologia desta pesquisa.

Ao meu amigo e colega, coronel e professor Marcos Antônio da Silva, com o qual mantive um rico e intenso intercâmbio de idéias sobre a pesquisa. Nessa mesma linha, ao meu aluno e amigo José Inácio Leiria, que recebeu os requisitos do protótipo por mim idealizado e que superou as expectativas do seu "cliente", ao construir um verdadeiro agente inteligente e ao fechar uma parceria muito promissora entre um trabalho de pesquisa de doutorado com as exigências do seu projeto de final de curso de bacharel em Sistemas de Informação. José Inácio: seu êxito na graduação será indubitável e sua continuidade na carreira acadêmica será indispensável. Enfim, aos meus colegas de mestrado e doutorado da UnB, aos meus alunos de graduação e aos meus companheiros engenheiros e topógrafos do Exército que, direta ou indiretamente, me ajudaram a concluir este curso.

Ao Criador dos seres e das coisas, pela existência, entre nós, do pesquisador britânico Tim Berners-Lee, atual presidente do Consórcio WWW e lançador dos fundamentos da rede mundial, a Internet, em 1984. Seu feito se igualou ou superou, na história, à invenção dos tipos móveis de João Gutenberg, há mais de meio milênio. Esta tese deve dividendos a essa privilegiada inteligência, cujo engenho de criação, de cunho *cosmocrático*, tem sido o motor potente para fomentar a criatividade pessoal e para promover as mudanças sociais que vêm alterando o equilíbrio do poder no mundo humano civilizado.

Finalmente, mais um preito ao Criador, por me presentear com uma família saudável e querida, com mais uma filha temporã no meio do curso e aos meus 44 anos de idade, com todas as oportunidades que me surgiram juntamente com as dificuldades, com a paciência da esposa, dos filhos e dos pais, que souberam muito bem o significado da palavra "ausência", pelo sem-número de pessoas maravilhosas que Ele pôs diante de mim nesse apertado prazo que me foi concedido para o curso e, por que não citar, as poucas pessoas, nada maravilhosas, que deliberada (por má-intenção e mesquinharia mesmo) ou involuntariamente (por ignorância) tentaram me prejudicar. Os senhores, as senhoras e vocês, ao lado das moléstias sem gravidade que afetaram a minha saúde e a dos meus entes queridos nesse período, ao lado até das ansiedades e incertezas de ordem psicológica que afligem normalmente um estudante de um curso desse nível, pois bem, esta pequena parcela de percalços serviu também de estímulo para a minha reação e para o meu sucesso. Muito obrigado, Senhor, por terdes me dado aquilo que os outros não querem, o que Vos restava em luta e em tormenta, representada por essas poucas pessoas e eventos inevitáveis, que me tornaram mais forte moral e intelectualmente!

RESUMO

BORGES, P.C.R. Avaliação da similaridade semântica entre classes de entidades espaciais, representadas por uma ontologia *ad-hoc*. Brasília: UnB. 2003. 384f. Tese (Doutorado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Faculdade de Estudos Sociais Aplicados da Universidade de Brasília.

A relação de similaridade semântica desempenha um importante papel nos sistemas de informação, por permitir a identificação e o tratamento de entidades ou de coisas que estão conceitualmente próximas. Dessa forma, tal relação constitui a base para projetar mecanismos de recuperação de informação que satisfaçam às necessidades dos usuários. Nos sistemas de informações geográficas (SIGs), até recentemente, os trabalhos de avaliação de similaridade entre entidades espaciais têm evidenciado tão-somente as características geométricas dessas entidades, as quais sempre recebem um tratamento isolado ou independente das relações semânticas subjacentes, com a finalidade de modelá-las em esquemas de bases de dados, que apresentam limitações em relação às propriedades cognitivas que a relação de similaridade semântica suscita, valendo citar a assimetria e a dependência contextual. O presente trabalho seguiu uma linha de pesquisa do Curso de Ciência da Informação Espacial da Universidade do Maine (EUA), cujo objetivo é estudar formas de inserir semântica em SIGs. As hipóteses que surgem nesta linha de pesquisa indicam que é possível (sob determinadas condições) simular num computador o senso humano de julgamento de similaridade de entidades espaciais. Na presente tese, numa primeira etapa, analisaram-se algumas tentativas de sobrepujar as limitações em relação às propriedades cognitivas do modelo vetorial. Numa segunda etapa, acolheram-se as concitações de pesquisa sugeridas por RODRÍGUEZ (2000), relacionadas à avaliação da similaridade semântica de entidades espaciais representáveis em bases de dados. Nada mais oportuno, de vez que, num período de trinta anos ou mais da recente história dos SIGs, grandes acervos de dados têm sido formados e muitos aplicativos de *geoprocessamento* têm sido desenvolvidos pelos enfoques tradicionais de tratamento da informação geográfica, de base geométrica, sem grandes compromissos de compatibilidade entre si (interoperabilidade) num nível semântico. E é justamente na interoperabilidade (em bases semânticas) que a linha de pesquisa seguida nesta tese é muito promissora, prevendo-se, num futuro próximo, a integração de dados oriundos de diferentes SIGs por agentes inteligentes. Para confirmar a hipótese de pesquisa, os instrumentos utilizados foram: um questionário e um protótipo de extensão ao modelo de simila-

ridade semântica de RODRÍGUEZ (2000), o MSS. A primeira versão deste protótipo teve o objetivo de auxiliar o entendimento do fenômeno da similaridade semântica e de gerar as variáveis do experimento. O modelo matemático desta primeira versão seguiu o enfoque vetorial de alguns autores. A finalidade da segunda versão do protótipo foi a de avaliar a similaridade semântica entre classes de entidades espaciais. O campo de observação bifurcou-se em duas unidades de observação: a primeira foi formada por entidades espaciais de uma região do sul do Brasil (Faxinal - PR), que foi mapeada por técnicas modernas de Engenharia Cartográfica. Estas entidades foram representadas numa base de dados orientada a objetos. A segunda unidade foi constituída por um grupo homogêneo de engenheiros e técnicos de Cartografia Automatizada, que foram submetidos a um questionário de avaliação sobre a similaridade entre as entidades citadas. A principal contribuição deste trabalho, além da já mencionada investigação sobre o estado atual das pesquisas congêneres sobre similaridade semântica e *ontologias*, foi estender o MSS para capacitá-lo a mensurar a similaridade semântica entre os objetos modelados numa base de dados espaciais. A base dessa avaliação não se refere às feições geométricas das entidades espaciais, mas aos termos (nomes) que as representam num subconjunto do modelo conceitual criado pela Diretoria de Serviço Geográfico do Exército Brasileiro (DSG) para a folha Faxinal.

ABSTRACT

BORGES, P.C.R. Assessment of semantic similarity among spatial entity classes represented by an *ad-hoc* ontology. Brasilia: UnB. 2003. 384f. Thesis (Doctor's degree in Information Science) – Information Science and Documentation Department of the Faculty of Applied Social Studies of the University of Brasilia.

Semantic similarity plays an important role in information systems by allowing the identification and management of entities or things that are conceptually close. As such, it constitutes a basis for designing mechanisms for information retrieval that satisfy users' needs. In geographic information systems (GIS), similarity assessment has so far been focused on geometric characteristics, which are treated in isolation and independent from the underlying semantic relations, in order to schematize them in a data base that are limited with respect to handling the cognitive properties of similarity, such as asymmetry and context dependence. This work has followed the research line of the Spatial Information Science Course of the University of Maine (USA), which goal is to study ways that could make GIS more semantic. The hypotheses that bear from this research line reveal that it is possible (under certain conditions) simulate the human sense of semantic similarity among entity classes in a computer system. In this thesis, at a first stage, it was drawn an analysis of some attempts to overcome the limitations concerning cognitive properties of the vector model. At a second stage of this work, some directions for future works from RODRÍGUEZ (2000) were followed, especially the one that recommends to assess semantic similarity among entity classes that could be represented in data bases. Nothing more well-timed to do in the recent history of GIS (the past thirty years or more) than this, inasmuch as huge data sets were gathered and traditional GIS softwares (geometry-based) were purchased, most of them is loosely committed with compatibility with each other (interoperability). And it is just on a semantic approach of interoperability that the research line followed in this thesis shows its strength and seems to be very promising, whereas being possible to foresee her proposal to be a first step to construct the future GISs, which will be managed by intelligent agents. In order to confirm the research hypothesis, two research tools were built: a questionnaire and a prototype to extend RODRÍGUEZ's (2000) semantic similarity model. The purpose of the first version of this prototype was to understand semantic similarity phenomenon and to help the researcher to derive the general variables. The mathematical model of this prototype was only based on a vector approach, according to other authors. The purpose of the second prototype was to

assess the semantic similarity among spatial entity classes. The observational field was divided into two observational units: the first one consists of spatial entities from a southern region of Brazil, called Faxinal (state of Parana). This area was surveyed with up-to-date cartographic engineering techniques and the spatial entities were represented in an object-oriented data base. The second one consists of a regular group of individuals: automatic cartography engineers (high level of education) and automatic cartography operators (medium level of education). The individuals have filled in a questionnaire in order to classify the entities according to their likeness. The main contribution of this work, besides the state-of-the-art on semantic similarity and *ontologies*, was the extension of the semantic similarity model to measure semantic similarity among the objects that are conceptually modeled in a spatial data base. The basis for this assessment is not related to the geometric features of the spatial entities, but to the terms (nouns) that represent them in a subset of the conceptual model created by the Directory for Geographic Service of the Brazilian Army (DSG) for the chart of the Faxinal region.

LISTA DE FIGURAS

LISTA DE QUADROS E TABELAS

LISTA DE APÊNDICES

LISTA DE EQUAÇÕES

LISTA DE DIREITOS AUTORAIS E PERMISSÕES DE USO

LISTA DE ABREVIATURAS E SIGLAS

1ª DL: Primeira Divisão de Levantamento
ABM: agente baseado em metas
AD: análise documentária
AG: algoritmo genético (plural: AGs)
AF: análise fatorial
AI: agente inteligente (pode ser um ARS, um AREI, um ABM ou um AU)
AREI: agente-reflexo com estado interno
ARS: agente-reflexo simples
AU: agente utilitarista
BC: base de conhecimento
BD: base de dados (plural: BDs)
CAD: projeto assistido (ou apoiado) por computador (*Computer-Aided Design*)
CASE: *Computer-Aided Software Engineering* ou Engenharia de Software Auxiliada por Computador.
CCAuEx: Centro de Cartografia Automatizada do Exército
CORBA©: *Common Object Broker Architecture* (Arquitetura de Agente de Requisição de Objetos, definida pelo OMG).
CSIS: sistema de informação de *software* compreensivo
CDOP: coleção de dados orientada a programa
C&T: Ciência e Tecnologia
CI: Ciência da Informação
CIGeo: Ciência da Informação Geográfica
DSG: Diretoria de Serviço Geográfico
EB: Exército Brasileiro
***fd*:** feição distintiva (plural: *fds*)
IBGE: Instituto Brasileiro de Geografia e Estatística
i.e.: isto é
IME: Instituto Militar de Engenharia

INPE: Instituto Nacional de Pesquisas Espaciais

IA: Inteligência Artificial

IG: informação geográfica

GPS: Sistema de Posicionamento Global (*Global Positioning System*)

KIF: *Knowledge Interchange Format*

LP: Lógica *Proposicional*

LPO: Lógica de Primeira Ordem

LRC: linguagem de representação do conhecimento (plural: LRCs)

LTP: linguagem técnica de programação (plural: LTPs)

MDT: modelo digital do terreno (plural: MDTs)

MMOO: métodos de modelagem orientada a objetos

MR: mundo-real

MSS: modelo de similaridade semântica

NCGIA*: National Centre for Geographical Information and Analysis*

OGC: *Open Geographic Information System Consortium*

OGIS: *Open Geospatial Interoperability Specifications*

OMG: *Object Management Group*

OO: orientado a objeto, orientação a objeto

OOAD™: *Object Oriented Analysis and Design*

OOSE™: *Object Oriented Software Engineering*

PCA: *principal component analysis*

PDI: processamento digital de imagens

p.ex.: por exemplo

pixel: elemento pictórico (*picture element*)

PLN: Processamento de Linguagem Natural

PMEGB: Projeto de Modelagem do Espaço Geográfico Brasileiro

POO: programação orientada a objetos

PROFAX: PROtótipo de avaliação da similaridade semântica para as classes de entidades espaciais da carta de **FAX**inal.

PRONTO©: PRotótipo de avaliação de similaridade semântica entre classes de entidades espaciais, representadas numa **ONTO**logia *ad-hoc*.

QEM: Quadro de Engenheiros Militares do EB

RB: rede bayesiana (plural: RBs)

RC: representação do conhecimento

RD: rede de Delaunay (plural: RDs)

RNA: rede neural artificial (plural: RNAs)

RS: rede semântica (plural: RSs)

SE: sistema especialista (plural: SEs)

SGBD: sistemas de gerenciamento de banco de dados

SI: sistema de informação (plural: SIs).

SIC: sistema de informações cartográficas (plural: SICs)

SIG: sistema de informações geográficas (plural: SIGs)

SIGAIA: sistema de informações geográficas apoiado na IA (plural: SIGAIAs)

simil: similaridade (variável)

SIS: sistema de informação de *software*

SPDOP: sistema de processamento de dados orientado para programa

SRI: sistema de recuperação de informação (plural: SRIs)

SS: similaridade semântica

TBCD ©: Tabelas da Base Cartográfica Digital (da DSG)

TI: tecnologia de informação (plural: TIs)

TMO ™: técnicas de modelagem de objetos (OMT™: *Object Modeling Technique*)

UML ™: linguagem de modelagem unificada (*Unified Modelling Language*)

V.: vide, veja.

vs.: *versus*.

SUMÁRIO

1. INTRODUÇÃO

***"Não há nada mais difícil de se tomar nas mãos, mais perigoso de se conduzir ou mais incerto quanto ao seu sucesso do que a iniciativa de se introduzir uma nova ordem de coisas."*[1] [Nicolau Maquiavel (1469-1527), estadista e historiador florentino]**

1.1. Generalidades

Supondo que uma pessoa esteja perdida numa esquina do centro de uma cidade, seria conveniente que ela tivesse acesso a um aparelho portátil como um celular ou um *palm-top*[2], dotados dos serviços WAP[2] e de um programa bem fácil de operar – interface gráfica -, que a levasse rápida e precisamente às informações que poderiam tirá-la desta situação. A interface disponível, hoje, na Internet, é baseada na pesquisa de palavras-chaves e texto. Se esta pessoa recorresse a um serviço de acesso dessa espécie (geralmente pago, como num *cybercafe*), procurando por informação sobre "mapas", e digitasse esta palavra no programa de busca da Internet, provavelmente o retorno seria de páginas e mais páginas de informações irrelevantes, até que ela achasse algo que lhe fosse conveniente.

O que é necessário nos dias de hoje, na Internet ou em qualquer serviço que recupere informações, é uma interface gráfica baseada no sentido da informação, no conteúdo das palavras. Os programas de pesquisa baseados no sentido das palavras e não apenas no texto são chamados de *pesquisadores semânticos*. É bom frisar que conter pesquisadores semânticos não é uma exigência apenas aplicada a aparelhos portáteis. É claro que suas dimensões acarretarão dificuldades adicionais em matéria de recursos de comunicação homem-máquina: espaço pequeno para exibição visual da pesquisa (tela) e teclado acanhado (sem muitas opções de teclas). Se estas dificuldades já são restritivas para recuperação de texto, é fácil imaginar como seria para a recuperação de mapas.

Quando se fala em mapas, neste exemplo, o apelo é para responder às indagações: "Onde estou? Para onde vou?". Trata-se de alguém que está perdido e que, essencialmente, deseja ser orientado sobre o local em que se encontra. Uma vez que o usuário dessa tecno-

---

[1] *Apud* PRESSMAN (1995).

[2] São computadores na forma de "caixinhas" de material plástico, com circuitos internos diminutos e que cabem na palma da mão. Outrora, eram considerados utensílios extravagantes e de utilidade duvidosa. Hoje, prometem ser uma ferramenta poderosíssima de recuperação de informação, juntamente com os celulares dotados de tecnologia WAP (*Wireless Application Protocol*), i.e., regras de comunicação (protocolo) que permitem o entendimento recíproco de 2 aparelhos portáteis que se conectam numa rede de longo alcance (Internet, p.ex.).

logia consiga comunicar à interface do seu aparelho que o assunto que ele deseja pesquisar está relacionado a lugares, a interface pode desligar centenas de opções de busca ligadas a assuntos médicos, financeiros, entre outros, concentrando-se basicamente em *localização* (FONSECA, 2000a) – eis aí uma forma de manifestação da *informação geográfica*.

Mas definido o assunto, a questão seguinte é: "Como a interface pode tratar essa informação?". Aqui, surge uma propriedade especial da informação geográfica – a parte gráfica -, que poderá influir numa formulação mais elaborada da pergunta precedente, desta forma: "Como relacionar os diversos níveis de representação da informação geográfica (também chamada de espacial) com os diversos níveis de pormenores semânticos, as palavras e seus sentidos?".

É normal, quando se está perdido (geograficamente ou não), começar a se procurar informação do nível mais alto (geral) para o mais baixo (pormenores). Por exemplo (continuando o anterior): se a pessoa perdida quiser um mapa da Av. Atlântica, na Zona Sul do Rio de Janeiro, é mais fácil começar com um mapa-múndi e apontar a seta de seleção do programa de busca para a América do Sul, depois para o Brasil, depois para a Região Sudeste, depois para o estado do RJ, depois, dentro deste estado, achar aquela grande mancha urbana que é a cidade do Rio de Janeiro; daí, ela começa a aprofundar a sua pesquisa, até chegar à localização da avenida do seu interesse (FONSECA, 2000a).

Há aparelhos portáteis conectados à Internet, mesmo desprovidos de GPS[3], que são capazes de fornecer uma localização aproximada da região em que estão. Isto pode facilitar a pesquisa, eliminando vários *cliques* em mapas até que se chegue ao nível de pormenor desejado. Poderia haver três opções de busca nesta interface: só informação espacial, só textual ou uma combinação dessas duas, o que seria muito conveniente para uma pessoa que possa ter achado o estado do RJ, mas que não soubesse onde se situa a sua capital. Como declarou FONSECA (2000a), uma tabela poderia ser juntada ao mapa, com o nome de algumas das principais metrópoles do RJ (cinco, por causa do tamanho da tela).

É importante que uma interface que tenha como objetivo ser direta ofereça poucas opções. Estudos da área cognitiva demonstram que os seres humanos se sentem melhor com uma faixa entre cinco e nove opções de cada vez. Mais que isso já é complicado, sobretudo com uma tela pequena como a de um celular.

Agora vem o ponto central deste caso hipotético, que diz respeito a esta pesquisa: como construir tal interface? Como fazê-la aceitar perguntas mais "semânticas" do usuário

---

[3] *Global Positioning System*: é um sistema de localização de qualquer ponto na superfície da Terra, com grande precisão, baseado no rastreio de uma constelação de mais de 20 satélites artificiais.

(qual a cidade ou o bairro?), em vez de perguntas muito minuciosas em termos de ângulos e distâncias, se o objetivo for recuperar informação geográfica? Como embutir nessa interface um conjunto de pontos de vista diferentes em relação ao usuário que a utiliza? Por exemplo, o enfoque da busca de informação pela interface deve adequar-se ao tipo do usuário que está perdido: se ele é um turista, procurando o Forte Copacabana, não pode ser-lhe apresentado um rol de respostas para um executivo que procura um restaurante para almoçar. O esforço de criar essa interface deve ser capaz de incorporar esses diferentes enfoques.

Trabalhar com mapas em equipamentos de mão é cada vez mais uma realidade. O que marcou essa passagem do sonho para a realidade foi o avanço da indústria na área da microeletrônica e das telecomunicações, levando a ciência a ser atropelada por sua grande tributária[4] – a tecnologia. Não obstante, cientistas vêm mantendo essa "corrida" sob certo controle. É o caso de um dos grupos de pesquisa da Universidade do Maine (EUA), na área científica recém-surgida no ambiente acadêmico norte-americano: a *Ciência da Informação Geográfica* (**CIGeo**)[5], especialmente preocupada com a inserção de semântica num sistema de informações geográficas (**SIG**)[6]. Esse gênero de sistema de informação, até hoje, tem sido modelado tão-somente em bases matemática (geométrica, essencialmente) e lógica. Em virtude disso, ocorrem efeitos adversos na interação do operador humano, sem profundos conhecimentos de Cartografia e de Informática, com um SIG, o que dificulta a comunicação homem-máquina.

Como resultado da conjunção de esforços técnico-científicos nessa área da CIGeo, surgem soluções inteligentes para máquinas pequenas, com limitados recursos de processamento e baixas taxas de transferência de dados. Estas soluções estão calcadas na personalização de acesso aos provedores de serviço de localização. Para isso, o provedor precisa "falar a mesma língua" do usuário. Esta forma de comunicação tem sido denominada no campo científico da Inteligência Artificial (IA) de ***ontologia***, termo tomado por empréstimo da Filosofia.

Lendo-se o artigo de FONSECA (2000a), percebe-se que as *ontologias* estão em alta no mundo dos sistemas de informação. Mas o que é ontologia? De uma tortuosa trajetória, que começou com Aristóteles (séc. IV A.C.), passando por Kant (séc. XVIII), a ontologia sempre foi definida de forma geral como sendo o tratado (*logos*) do ser (*ontos*), independente de como ele se manifeste. O termo acabou sendo apropriado pela IA e ganhou uma defi-

[4] Tradicionalmente, a recíproca sempre foi verdadeira: a tecnologia sempre se beneficiou dos avanços científicos.
[5] Não seria conveniente uma sigla como CIG, pois haveria confusão com SIG, há mais tempo consagrada na comunidade científica brasileira como sistema de informações geográficas.
[6] No anglo-americano é conhecido como **GIS** ("Geographic Information System"). V. glossário.

nição mais palpável. *Grosso modo*, ontologia representa a visão do mundo pelo enfoque de uma determinada comunidade, contendo todos os conceitos e relações entre os conceitos que são do interesse dessa comunidade. Portanto, diferentemente do conceito filosófico, não existe apenas uma ontologia - a ontologia. Para os engenheiros de *software* e cientistas da computação, existem tantas ontologias quantas forem as visões (enfoques) das comunidades de usuários.

Vê-se que a definição de ontologia está estreitamente ligada ao conceito de linguagem e que a Lingüística terá uma papel relevante nos estudos interdisciplinares que surgirem.

Voltando aos serviços de localização de informação geográfica, as ontologias vão funcionar como estruturas integradoras de informação. Seria o Santo Graal para os cientistas e engenheiros que labutam há mais de trinta anos em SIG, para chegar a um padrão de compatibilidade[7] universal para esses sistemas. Nesse caso, uma ontologia funcionaria de maneira quase inversa, assumindo dois papéis: por um lado, como estrutura concatenadora de informações de diversas fontes; por outro, como disseminadora seletiva e coerente de informações pelas comunidades de usuários que reclamam por informações.

Intenta-se, com esta pesquisa, contribuir com os esforços de construção de SIGs *interoperáveis*, dotados de recursos que utilizem ontologias dedicadas ao usuário de informação geográfica. Além disso, almeja-se somar esforços às pesquisas realizadas no Centro Nacional de Informação Geográfica e Análise[8] da Universidade do Maine, em que uma série de trabalhos têm mantido uma linha de continuidade para alcançar a meta de *interoperabilidade* já citada.

Uma das linhas de pesquisa na área de sistemas de informação é a que lida com **similaridade semântica**. No caso de SIGs, a similaridade semântica está ocupada com a identificação e o tratamento de objetos espaciais que estão conceitualmente muito próximos, não se preocupando com as clássicas formalidades da Matemática e da Lógica, que, sozinhas, não propiciaram uma forte interação na comunicação homem-máquina. Os aspectos semânticos, de natureza mais subjetiva (área cognitiva), facilitam mais esta comunicação, apesar de serem difíceis de modelar e tratar em sistemas automatizados de informação, de características formais.

---

[7] No jargão do pessoal da área de sistemas de informação: *interoperabilidade.*
[8] **NCGIA**: *National Center for Geographic Information and Analysis*

1.1.1. Organização do trabalho de pesquisa

Este trabalho foi organizado em onze seções, das quais sete constituem capítulos dedicados à pesquisa propriamente dita. As outras quatro seções são porções acessórias da tese, que tratam das referências bibliográficas (item 8), apêndices (item 9), relação de direitos de autoria e de permissão de uso de produtos digitais (item 10) e um glossário de termos das geociências e das ciências cognitivas (item 11).

Por conta e risco do pesquisador, foi ampliado em extensão o capítulo introdutório da tese, em virtude da sua natureza exploratória, a fim de aclimar os leitores dos diversos campos do conhecimento que o trabalho envolve. O preâmbulo conceitual que se inicia no subitem 1.2 e vai até o 1.5 contém o levantamento bibliográfico que norteou a fase mais pormenorizada da revisão de literatura e não deixa de contribuir com um objetivo tácito desta pesquisa, que foi o de ajudar a levantar o estado-da-arte sobre a similaridade semântica no campo do *Geoprocessamento*.

A organização básica do texto tomou por regra introduzir os tópicos tradicionais de um trabalho científico nos seus aspectos gerais e, mais adiante, tendo por interface a revisão de literatura, voltar aos mesmos tópicos, mas de forma mais minuciosa. Está aí a razão para se apresentar o problema e o objetivo geral de pesquisa no Capítulo 2, realizar a revisão de literatura no Capítulo 3, para, então, derivar os problemas e objetivos específicos, assim como as hipóteses estatísticas, no Capítulo 5.

A hipótese de pesquisa, nas suas formas alternativa e de nulidade, não foram introduzidas juntamente com o problema e objetivo geral de pesquisa no Capítulo 2, em razão da precocidade dos conhecimentos do pesquisador naquela altura dos estudos.

Em que pese ter sido realizada uma antecipação da formulação da hipótese de pesquisa por meio de uma análise racional (*rationale*) no subitem 2.3.1, a finalização segura do seu enunciado só se deu no Capítulo 4, num ponto da pesquisa de maior maturidade sobre os conhecimentos adquiridos para a formulação deste tópico-guia para a extração dos objetivos específicos e para o delineamento da metodologia desse estudo-de-caso exploratório, de manipulação experimental.

Os Capítulos 3 e 6, por se tratarem de partes estreitamente ligadas por relações de manancial e sumidouro de informações, respectivamente, compartilharam uma característica: sínteses ou conclusões parciais ao término de cada subitem principal (primeiras subdivisões dos capítulos), que serviram para orientar a consolidação das conclusões finais do Capítulo 7 e até mesmo para justificar as concitações a trabalhos futuros deste capítulo, o que constituiu um valioso artifício de racionalização e de organização de trabalho numa pesquisa de

cunho exploratório, caracterizada pelo largo espectro de conclusões e de problemas para indicar em pesquisas futuras.

Fugindo um pouco das especificações de forma para trabalhos desse gênero, foi incluído um item (10), para abonar a origem dos aplicativos (*software*) que, ou foram usados diretamente nesta pesquisa, ou foram nela produzidos, ou foram exaustivamente citados ao longo do texto. Esta inserção poderia consumar-se num anexo, mas se optou por estender a lógica de especificação para uma relação de itens como figuras e tabelas, por exemplo, a fim de acentuar o papel que esses itens (cedidos, comprados ou produzidos) tiveram na tese.

Outra característica de uma pesquisa exploratória que opera numa área de limbo entre dois[9] campos científicos como esta, é a profusão de conceitos ainda fora do alcance da maior parte do público-alvo (cientistas da informação, cientistas da computação e engenheiros cartógrafos, particularmente os lusófonos). Para cobrir possíveis lacunas de entendimento, julgou-se coerente incluir um glossário de termos de ambos os campos em questão. No entanto, para não avolumar o tomo do trabalho, que provavelmente acabaria por ser bipartido, se fosse incluído o glossário em base celulósica (papel), optou-se por transformá-lo num arquivo digital.

O suporte magnético utilizado para armazenar o arquivo acima mencionado foi um CD-ROM, que também contém as tabelas do SPSS™ de consolidação das respostas às seis perguntas do questionário, o código-fonte em *Java*™ e todos os arquivos necessários para a execução do PRONTO©, em qualquer sistema computacional compatível com a plataforma *Windows*® (9x e posteriores). Este CD está acondicionado num invólucro apropriado, colado à contracapa final do tomo da tese.

1.2. Antecedentes

Nos três subitens seguintes são expostos os fatos e atos de natureza econômica, científico-tecnológica e até mesmo psicossocial, que afetaram direta ou indiretamente a eclosão das TIs de *geoprocessamento*, incluindo-se aí, especialmente, os sistemas de informações geográficas (SIGs).

Apesar de parecerem redundantes os subitens 1.2.1 (*geociências*) e 1.2.3 (Ciência da Informação Geográfica - CIGeo), o primeiro situa a produção e a difusão da informação geográfica num contexto mais abrangente, de ordem econômica e científico-tecnológica, além de explorar longitudinalmente o problema no relativamente curto período de gênese dessas

TIs. O segundo pende mais para aspectos de ordem acadêmica e psicológica dos pensadores das *geociências* e dos usuários das tecnologias derivadas de *geoprocessamento*, ao focar nos fatos e atos mais recentes, ligados ao surgimento da CIGeo no Brasil.

1.2.1. No campo das *geociências*

Segundo STAR (1990), o primeiro sistema que surgiu para tratar a informação geográfica, num ambiente moderno de microeletrônica e informática, foi o **CGIS** ("Canada Geographic Information System"), em 1964, concebido para incrementar as políticas de cadastro urbano e rural canadenses e avaliar a interferência antrópica no meio-ambiente. Quando surgiu este sistema, para auxiliar o governo canadense na resolução dos problemas citados, não se podia ainda vislumbrar o papel que tal tecnologia desempenharia no apoio, complementação e interação dos mais díspares campos do conhecimento humano, em pouco mais de trinta anos.

A característica multidisciplinar de um SIG é a fonte de seu sucesso, não se restringindo a sua aplicação apenas às áreas específicas das chamadas *geociências* (Cartografia, Geologia, etc.). Na verdade, tem sido crescente o emprego dessa poderosa ferramenta em grandes empresas governamentais e privadas como peça fundamental de apoio à decisão e para a análise dos resultados obtidos das mais variadas formas de cruzamento de informações textuais e gráficas.

Problemas críticos da vida quotidiana do homem, como os ligados ao meio-ambiente, quando são razoavelmente resolvidos por tecnologia viável em termos econômicos e cronológicos, creditam indubitável sucesso aos produtos e fortuna aos seus produtores. Estas tecnologias[10] de aquisição, processamento e análise de dados cartográficos digitais[11], seguindo-se à posterior exibição da informação geográfica, nas mais diversas formas, logo se mostraram muito atraentes comercialmente e logo surgiram as grandes indústrias de programas e de equipamentos especializados nas diversas fases do processo de produção de documentos cartográficos (mapas e cartas topográficas). Cada uma das grandes indústrias "enxergou" e ainda vem "enxergando" o mundo empírico dos fatos geográficos ao seu próprio modo, sem embasamento conceitual suficiente e, desta forma, todas vêm produzindo sistemas que não "se entendem" entre si.

---

[9] *Geoprocessamento* e ciências cognitivas (incluída a Ciência da Informação).

[10] Hoje, designadas pelo termo *Geoprocessamento* (V. glossário).

[11] É bom impor mais rigor na distinção entre dado cartográfico e informação geográfica. Dados cartográficos estão intimamente ligados ao método de engenharia (cartográfica, no caso), que é quantitativo por excelência. Quando se está no âmbi-

A partir de 1964, os interesses comerciais, a reboque da inovação tecnológica, subverteram todo o trabalho de fundamentação teórica dos SIGs. Os estudos teóricos não acompanhavam o ritmo da produção em série dessas ferramentas. Somente em meados da década de 90, quando uma parcela significativa de usuários já havia investido soma vultosa de recursos na compra de *software* e na montagem de bases de dados cartográficos, é que as pesquisas sistematizadas começaram a identificar os efeitos colaterais adversos dessa tecnologia, em confronto com uma outra tecnologia da informação que tomava forte impulso – as redes.

E qual seria o "calcanhar-de-aquiles" da primeira tecnologia? O ponto fraco da primeira é agravado exatamente quando ela se conjuga com a segunda, i.e., quando opera num ambiente distribuído de dados. Trata-se do intercâmbio, da troca de dados cartográficos numa forma consistente, independente das limitações de mecanismos de acesso, de arquiteturas de *hardware* e de estruturas específicas de *software* dos sistemas comerciais. Esta problemática permanece subjacente no presente trabalho de pesquisa e denomina-se *interoperabilidade de sistemas de informação geográfica.*

É instrutivo investigar a causa da diversidade de enfoques das grandes indústrias de tecnologias de informação geográfica. Esta diversidade e as eventuais confusões no intercâmbio de dados entre os grandes sistemas não ocorreram com o lançamento da tecnologia dos sistemas de bancos de dados relacionais, cuja comunidade de usuários e desenvolvedores criou um acervo considerável de produtos confiáveis e até hoje produtivos, graças à sólida teoria criada por Edgard F. Codd, pesquisador da IBM, em 1970.

A ambigüidade que afeta a informação geográfica não existe porque os usuários e organizações que desenvolvem esses sistemas assim o desejam, mas em função de a tecnologia ter sido implementada antes de se ter estabelecido um arcabouço teórico sobre a natureza da informação espacial (CÂMARA, 1998).

Eis aí a fonte de todos os problemas atuais de recuperação, integração e manutenção da informação geográfica. Isto prejudica sobremodo a interoperabilidade entre esses sistemas, já que a tecnologia adiantou-se sobremodo aos estudos e ao estabelecimento de um arcabouço teórico sobre a natureza e as relações conceituais que envolvem a informação geográfica.

---

to da Geografia, sobe-se no nível de abstração e é mais preciso empregar-se informação geográfica (nível de síntese dos dados cartográficos), mais susceptível aos métodos qualitativos das ciências sociais.

1.2.2. No campo das ciências cognitivas

O termo "ciências cognitivas" enseja um sem-fim de discussões polêmicas de bases filosófica e psicológica, que fogem ao objetivo desta tese investigar a fundo, sendo que é um estudo-de-caso exploratório que se apóia em alguns princípios da Inteligência Artificial, disciplina da Ciência da Computação. Todavia, uma breve revisão sobre as bases da Psicologia Cognitiva, "berçário das ciências cognitivas", ajudará a projetar o cenário sobre o qual surgiram os marcos teóricos que nortearam a pesquisa. Alguns outros aspectos sobre as discussões polêmicas que se encaixam no tema constam no subitem 1.5.4.

Esta pesquisa explora assuntos ligados a meios de comunicação[12] como mapas e outros documentos cartográficos [(BERTIN, 1967), ANDRÉ (1980), (MARTINELLI, 1991), (MOURA, 1994), (PRATT, 2000) e (PRADO, 2001)], sujeitos à interação com outras manifestações de comportamento de seres inteligentes (representação, interpretação e memória, p.ex.), sendo necessário remontar, ainda que de forma sucinta, de onde surgiu a noção de similaridade semântica, que sobressai no próprio título do trabalho.

Segundo KELLER (1964), tudo começou quando *René Descartes*, em meados do séc. XVII, lançou a sua famosa distinção entre *corpo* e *mente* e abriu duas correntes de idéias que ainda se confrontam, mas que, ao mesmo tempo, se enriquecem no âmbito da Psicologia moderna. As três principais tributárias desta última começaram a surgir no final do século XIX:

- ***Estruturalismo***: os seus primórdios situam-se nos estudos sobre a visão humana, com o médico alemão *Herman von Hemholz* (1832-1920), continuados por Guilherme Wundt, seu aluno. Em 1879, Wundt, na Universidade de Leipzig, introduziu a metodologia experimental nos estudos psicológicos, influenciando o seu aluno E. B. Titchener (1867-1927), da Universidade de Cornell, considerado o fundador desta escola nos EUA;
- **Funcionalismo**: escola considerada como a principal fonte de influência para a Psicologia Cognitiva (EYSENCK, 1994). Surgiu em 1890 com a publicação da obra "Principles of Psychology", de William James. Escola muito influenciada pelo *darwinismo* (comportamento adaptativo), teve como pedra angular da sua muito criticada teoria valorizar mais a função do que a estrutura cerebral, ampliando o campo de estudo psicológico para o estudo do comportamento animal, infantil e anormal (KELLER, 1964). Na verdade, é uma corrente da Filosofia da Mente, cuja tese é a de que não só o cérebro de um ser biológico inteligente é capaz de criar instâncias (estados) de processos cognitivos, mas que estruturas não-biológicas também são capazes de "possuir determinados tipos de mente",

p.ex: circuitos integrados em pastilhas (*chips*) de Si (silício) ou Ge (germânio), como nos microprocessadores (ABRANTES, 1994);

- ***Comportamentalismo* (*behaviorismo*)**: muito ligado ao aspecto corporal das idéias de Descartes, surgiu em 1919 com a publicação da obra intitulada "Psychology from the standpoint of a behaviorist", de John B. Watson (1878-1958), na Universidade de Chicago (EUA). O centro de preocupação desta escola estava na observação objetiva do comportamento, que é reduzido ao binômio estímulo – resposta (produzida por músculos e glândulas). ABRANTES (1994) assevera que boa parte das críticas ao funcionalismo vêm desta escola, que considera inacessíveis aos métodos científicos os processos mentais de "aprendizagem", "representação", "crenças", "conhecimento", etc.

Além das três escolas vistas acima, havia mais três, que também contribuíram de alguma forma para a chamada Psicologia moderna: o *gestaltismo*, o *finalismo* (ou escola *hórmica*) e a *psicanálise*. No entanto, de todas as escolas, foi o *behaviorismo* que mais traços passou para a Psicologia moderna (KELLER, 1964).

Mas não é na Psicologia que se assenta o marco teórico desta pesquisa. É nas conseqüências das pesquisas na Psicologia Cognitiva, área do conhecimento já considerada por muitos autores (EYSENCK, 1994)[13] como independente do tronco da Psicologia. Nessa área já se vem consolidando uma unidade metodológica autônoma, que muito deve à escola funcionalista. Dentro da Psicologia Cognitiva, três disciplinas começam a se definir: a Psicologia Cognitiva Experimental, a Neuropsicologia Cognitiva e a Ciência Cognitiva[14].

Em breves palavras, os *psicólogos cognitivos experimentais* tratam de sujeitos normais por meio da pesquisa empírica tradicional, sem se utilizarem da metáfora computacional[15].

Os *neuropsicólogos cognitivos* investigam os padrões de *deficit* cognitivo apresentados por pacientes com lesão cerebral e os relacionam ao funcionamento normal.

Os *cientistas cognitivos* combinam a pesquisa empírica com a modelagem computacional de problemas de cognição humana, emprestando um rico material metodológico e conceitual para a Inteligência Artificial. A metáfora computacional é fundamental para um cientista cognitivo. Mais sobre metáforas (científicas), como formas alternativas de desvendar conhecimento, será visto no subitem 1.2.3.

---

[12] Linguagem.

[13] Estes autores apelaram para o paradigma da revolução científica e o conceito de ciência pós-moderna de Thomas S. Kuhn ("A Estrutura das Revoluções Científicas" - 1962).

[14] Maiores informações sobre estes campos científicos, no glossário.

[15] Importação terminológica intensa da Ciência da Computação e da Teoria da Informação (LÉVY, 1993).

Por outro lado, alguns autores (ABRANTES, 1994), mais cautelosamente, admitem a existência de um núcleo ainda multidisciplinar voltado para o estudo da mente, ao qual denominam de "ciências cognitivas", formado pela Lingüística, pela Ciência da Computação (IA), pela Epistemologia e pela Neurofisiologia.

Pode-se dizer que o ano de 1956 foi o marco-zero para a Psicologia Cognitiva, como sucessora e destituidora do *behaviorismo*. Fatores surgidos em meados do séc. XX, como o novo enfoque para definir ciência e o surgimento do computador eletrônico, foram decisivos para isso, por exemplo:

- O trabalho de Kenneth Craick ("A Natureza da Explicação"), de 1943, que introduziu um passo intermediário ao modelo *behaviorista* estímulo - resposta, denominado *etapa de processamento cognitivo* (raciocínio, crenças, propósitos), com a finalidade de explicar o comportamento humano (RUSSELL, 1995);
- Claude E. Shannon e W. Weaver, em 1949, publicaram o artigo "Teoria Matemática da Comunicação"[16], ante-sala para a revolução da microeletrônica e da Informática. Este trabalho foi uma extensão da tese de mestrado do primeiro autor, no MIT, em 1938, em que aplicou a Álgebra de Boole à análise de circuitos de relés, (DAGHLIAN, 1995);
- A famosa Conferência de Dartmouth (Hanover, N. Hampshire), no verão de 1956, na qual participaram John McCarthy, Marvin L. Minsky, Noam Chomsky (apresentou um estudo preliminar da sua Gramática Transformacional da Linguagem), Allen Newell e Herbert A. Simon (discutiram o modelo computacional do seu Solucionador Geral de Problemas), entre outros, quase todos virtuais fundadores da IA (EYSENCK, 1994).

Até a década de 60, o *behaviorismo* ainda era a escola dominante nos EUA. Psicólogos como Kenneth Craick e outros que trabalhavam com noções de *conhecimento*, *crenças*, *propósitos*, *mente*, *etapas do processo de raciocínio*, *agente inteligente* (AI) e outras congêneres eram considerados como *psicólogos populares* pelos *behavioristas*.

Apesar de lançar teorias sustentáveis sobre o comportamento animal de certas espécies, o *behaviorismo* não foi tão bem-sucedido quando se tentava usá-lo para explicar o comportamento inteligente de outras espécies animais de ordem superior como a humana.

O plano básico desses *psicólogos populares* (futuros cientistas cognitivos) para um AI (agente inteligente) era o seguinte: 1) O agente já deveria trazer em si um modelo em pequena escala do mundo (realidade); 2) Este modelo deveria encerrar um conjunto próprio de regras de ação capazes de ser empregadas segundo as alternativas (representações) que seus processos cognitivos estruturassem; 3) A seguir, havendo mais de uma alternativa (re-

presentação) produzida, o agente deveria decidir qual seria a melhor a ser tomada (RUSSELL, 1995). O fundamento da melhor decisão do AI vem do princípio da racionalidade, cujas origens remontam o séc. IV A.C., com a ética do pós-socrático Epicuro: "Ser racional é fazer a coisa certa[17]"; 4) Sistema de armazenamento (memória) e de recuperação de informações para uso em situações futuras, similares às já experimentadas (aprendizagem).

Já na década de 70, os psicólogos cognitivos começaram a discutir a aplicação do enfoque da Teoria da Informação aos complexos processos da cognição humana. Foram detectadas falhas nesta aplicação, porque as diferenças individuais não eram contempladas pela teoria e porque era preciso restringir muito o modelo experimental, já que a teoria considera o sistema cognitivo isolado de influências motivacionais e emocionais.

A IA, virtual herdeira da bagagem teórica do funcionalismo[18], não poderia ficar imune a essas controvérsias. Boa parte dos psicólogos atuais, como já aventado, ainda muito ligados aos métodos *behavioristas*, não concorda com o estabelecimento de uma IA como campo científico independente, justamente porque, na IA, se defende o enfoque computacional da mente (base de seu problema central). Este enfoque computacional, por sua vez, baseia-se na manipulação de símbolos segundo regras ou procedimentos (algoritmos) sensíveis, unicamente, às propriedades desses símbolos (sua forma), independentemente dos significados associados a tais símbolos. Por esse enfoque foram construídas diversas linguagens de programação que controlam as operações de máquina de maneira predominantemente sintática.

Como se não bastassem as críticas externas à IA, dentro dela própria há desentendimentos sobre que modelos instanciar os processos mentais. Daí surgiram duas linhas de pesquisa: os tradicionais ***cognitivistas***, que defendem linguagens lógicas para descrever esses processos e os ***conexionistas***, que formam uma espécie de ponte entre os seus colegas cognitivistas e os psicólogos modernos, já que constroem modelos que se inspiram na Neurofisiologia.

Os *cognitivistas* ficam satisfeitos com o tradicional modelo seqüencial da computação, o Ciclo[19] de von Neumann: aquisição – decodificação – execução, de dimensão essencialmente sintática, inadequado à representação de muitos processos cognitivos. O que importa para eles não é a arquitetura computacional (baixo nível, nível de máquina ou subsimbólico),

---

[16] Teoria da Informação (RUSSELL, 1995).

[17] O "certo", depois de passar pelos materialistas do séc. XVII (Tomás Hobbes), pelos utilitaristas e pelos pragmáticos dos séc. XVIII e XIX (Jeremias Bentham e Stuart Mill), passou a ter uma conotação mais objetiva na forma "adequado".

[18] Segundo vários autores da Psicologia Cognitiva e da IA: EYSENCK (1994), RUSSELL (1995), etc.

[19] Também: Máquina de von Neumann.

mas o poder de expressividade dos programas, i.e., a capacidade de um algoritmo expresso em declarações de uma determinada linguagem técnica de programação (LTP) processar regras, calcular resultados e mesmo derivar novas regras para uma base de conhecimentos inicialmente carregada num AI. Em suma, o modelo de computação tradicional é suficiente para incorporar o simbolismo da LPO (Lógica de Primeira Ordem), que é a forma de representação do conhecimento mais adequada ao formalismo lógico-matemático.

Os *conexionistas*, por outro lado, reclamam do modelo tradicional por achar que ele já atingiu o seu limite físico (PASSOS, 1990). Acreditam os conexionistas que o calcanhar-de-aquiles cognitivista está justamente em desprezar os pormenores da arquitetura do sistema computacional, porque é o nível mais básico da máquina que impõe restrições importantes aos mais altos, podendo criar graves empecilhos para os construtores de programas nesses níveis mais elevados. Daí, concluem os conexionistas, se forem contornados tais empecilhos, abre-se um espaço de inusitadas soluções para problemas apresentados pelos maiores críticos[20] à IA, particularmente aqueles que alegaram e até demonstraram que a IA não poderia ser considerada um campo científico por não poder descrever e predizer muitos dos complexos processos do comportamento racional, especialmente a linguagem.

O conexionismo parece prometer agregar a dimensão semântica aos seus modelos de uma forma mais natural, o que fica bem mais difícil para um modelo da vertente cognitivista. Essa dimensão semântica, no conexionismo, está estreitamente relacionada ao estado geral de uma rede de elementos de processamento (EPs) que simulam os neurônios humanos.

O conexionismo propõe uma arquitetura alternativa ao modelo tradicional de computação, denominando-a de Computação Neural (PASSOS, 1990) ou ainda: Neurocomputação, Processamento Paralelo e Distribuído (PDP), Sistemas Neuromórficos, Redes Neurais Artificiais (RNAs) ou Sistemas Neurais Artificiais (SNAs), termos muitas vezes usados como sinônimos no âmbito da Ciência da Computação, segundo GONDIM (1991). Esse modelo difere primordialmente do tradicional pelas seguintes características, que se assemelham bastante a muitas das características dos processos ligados ao comportamento humano inteligente:

- Tolerância a falhas de funcionamento (inadmissível num modelo baseado na LPO);
- Não é seqüencial e nem sempre é determinístico;
- Incontestável paralelismo temporal no processamento da informação;
- Capacidade de aprendizado pela experiência;

---

[20] Os da linha do filósofo John Searle e da linha do lingüista Noam Chomsky, p.ex.

- Alto grau de conexão (sinapses) entre seus EPs;
- Adaptabilidade.

O confronto atual de idéias nessas áreas que emergem é o problema central da Filosofia da Ciência, ocupada com a demarcação do que é e do que não é ciência, fugindo aos compromissos traçados neste projeto de pesquisa. Pelo escopo a ser definido para esta pesquisa, é impossível discorrer a fundo sobre conceitos ligados à natureza da inteligência, significado, língua, linguagem e outros dessa ordem. A revisão de literatura tratará, em linhas gerais, de conceitos que sustentem o objetivo da pesquisa, fugindo da polêmica. Sem embargo, esse breve apanhado de definições e mesmo de reflexões ajudará a colocar o leitor alheio a esses campos num ponto de relativa ambientação com o estado atual da IA, que de alguma forma, direta ou não, induziu diversos trabalhos que procuraram tratar da quantificação de fenômenos de natureza cognitiva.

A próxima seção ilustra rapidamente a recentíssima história dos SIGs orientados à cognição. De certa forma, é uma seção que mostra a convergência dos sistemas tratados no subitem 1.2.1 para as soluções de natureza cognitiva tratadas no subitem 1.2.2.

1.2.3. No campo da Ciência da Informação Geográfica

Segundo a literatura a que se teve acesso, parece que as pesquisas aplicadas que povoam áreas de interesse comum entre as ciências cognitivas e as *geociências* começaram na década de 90.

Os conceitos dos trabalhos na região de limbo entre a Lingüística e a Computação não passaram despercebidos pelos cientistas da área *geocientífica*. KUHN (1993), cientista da área de SIG, foi referenciado no artigo de BÄHR (1996) pelo seu trabalho intitulado *"Metaphors Create Theories of Users"*, cujo teor resume-se à confirmação do fato de que os seres humanos, nas suas constantes relações, utilizando-se primordialmente da língua natural, preferem também enriquecê-la ou complementá-la com outras formas de linguagem que não sejam as mais formais da Matemática e até mesmo as mais descritivas como a própria representação escrita da língua natural. É intuitivo que as pessoas prefiram um **esboço** ou rascunho a um gráfico ou diagrama muito minuciosos; que prefiram a **escrita** na descrição de um fato a equações complexas da Álgebra; e que prefiram **gestos** ou a fala fluente à própria escrita.

LOPES (1987) e ORTONY (1988), na segunda metade da década de 80, já traziam à tona a chamada "metáfora científica", extrapolando as aplicações estilísticas tradicionais, em

que se substitui o sentido natural de uma palavra por outro sentido, em virtude de uma comparação não enunciada.

A metáfora sempre fora reduzida a uma matéria subsidiária da literatura. Nessas obras, contudo, os autores concordam em que a metáfora produz um "conhecimento revelado", que freqüentemente desafia os conhecimentos científicos construídos pela razão.

São vários os exemplos que denotam a força da metáfora como forma de incrementar a imaginação criadora do cientista ou do artista. Nas invenções, esses exemplos são freqüentes, citando-se Guilherme Harvey, médico inglês do séc. XVII, que, num momento de *inspiração (metafórica)*, associou os movimentos mecânicos de uma bomba aos da circulação sangüínea, estabelecendo as bases teóricas desse fenômeno fisiológico. É o resgate da força cognitiva desse instrumento de sondagem do raciocínio científico mais abstrato (LOPES, 1987).

A IA sai ganhando com a metáfora, visto que uma das contribuições oriundas da Lingüística Computacional deu origem a um dos ramos de pesquisa aplicada da IA: o **PLN** (Processamento de Linguagem Natural).

Outros avanços desafiam os produtores de *hardware* e de *software*. O que BÄHR (1996) prognosticara em meados da década de 90 começou a tornar-se problema de pesquisa no início do fluente século, p.ex., fazer do sistema computacional um auxiliar no processo de visualização de um problema complexo.

Em BLASER *et al.* (2002), p.ex., o entendimento do processo de visualização é um estágio preambular para sistemas especialistas com capacidade de resolver problemas de SIG. Os autores abandonaram as definições limitadas do termo e partiram para um enfoque multidimensional, em que a visualização está ligada não só à mera representação gráfica de objetos numa tela do monitor de vídeo de um computador, mas também à construção de uma *imagem mental* desses objetos.

Esse enfoque implica interação entre o usuário e o sistema computacional, de tal sorte que ambos colimem um só objetivo: a máquina (o sistema) deve ser capaz de capturar e interpretar estímulos oriundos de um usuário humano e que este não necessite de apurados conhecimentos tecnológicos para submeter as suas necessidades à máquina. Apesar de não ser o foco desta pesquisa, percebe-se que esse é um típico caso de comunicação, envolvendo problemas de interface e linguagem, já imaginados por WILLIAMS *et al.* (1995).

Em nenhum outro momento da recente história das tecnologias da informação, a metáfora computacional foi levada a tão extrema realização, saindo das bases teóricas, muitas vezes controversas, para atingir níveis objetivos, em que os resultados justificam até a cria-

ção de linhas de pesquisa em cursos de Engenharia da Computação e que acabam por formar campos autônomos e institucionalizados do conhecimento, como é o caso do **CCIGeo** (Curso de Ciência da Informação Geográfica), na Universidade do Maine (EUA), especialmente preocupado com a inserção de semântica num sistema de informações geográficas.

As bases teóricas que às vezes suscitam polêmica têm sido acompanhadas de perto pelos *geocientistas*, a fim de que não se repitam os erros do passado, quando a tecnologia era divulgada e notadamente comercializada sem a necessária sustentação teórica, que só a ciência garante (HUISMAN, 1976). Foi desta falta de sinergia entre ciência e tecnologia que sobrevieram todos os problemas de interoperabilidade entre SIGs, até hoje existentes.

Daí, têm surgido tentativas transdisciplinares para evitar a dissociação entre os sistemas de processamento de informação geográfica do século XXI.

CÂMARA (2002), p.ex., sistematizou os mananciais teóricos, classificando em 4 grandes ramos as teorias sobre a informação geográfica (IG) e o *Geoprocessamento*. No Quadro 1.1, para cada escola, é mostrado o conceito-chave em sua definição de espaço, a representação computacional que melhor aproxima este conceito e algumas técnicas típicas de análise geográfica, que estão associadas a essa escola geográfica.

Para salientar as principais características de cada escola exposta no Quadro 1.1, segue-se um resumo elucidativo:

- **Geografia *Idiográfica*** (SIGs da década de 80): o conceito-chave é a unicidade da região, expresso por abstrações como a *unidade de área*[21]. A representação computacional associada é um polígono, com seus atributos geralmente expressos numa tabela de um banco de dados relacional e as técnicas de análise comuns, por meio da interseção de conjuntos (Lógica Booleana).
- **Geografia Quantitativa** (SIGs de hoje): o conceito-chave é a distribuição espacial do fenômeno de estudo, expressa por intermédio de um conjunto de eventos, amostras pontuais ou dados agregados por área. A representação computacional associada é a superfície (expressa como uma grade regular) e há uma grande ênfase no uso de técnicas de Lógica Nebulosa (*Fuzzy*) para caracterizar as distribuições espaciais.
- **Geografia Quantitativa** (SIGs da próxima geração): o conceito-chave são os modelos prospectivos (*preditivos*), com representação espaço-temporal, em que a evolução do fenômeno é expressa por uma representação funcional. Para capturar as diferentes relações dinâmicas, as técnicas de análise deverão incluir modelos multiescalares, que esta-

---

[21] *Constructo* classificador de partes de um certo espaço geográfico a ser carregado num SIG.

beleçam conexões entre fenômenos em pequenas escalas (tipicamente relacionados com fatores econômicos) e fenômenos em grandes escalas (tipicamente associados ao uso da terra).

Quadro 1.1 Teoria Geográfica e *Geoprocessamento*[22]

Teoria	Tecnologia GIS associada	Conceito-Chave	Repres. Comput.	Técnicas Análise
Geografia Idiográfica	Anos 80 – meados dos anos 90	Unicidade da Região (unidade-área)	Polígono e atributos	Interseção conjuntos
Geografia Quantitativa-1	Final da década de 90	Distribuição Espacial	Superfícies (grades)	Geoestatística + lógica "fuzzy"
Geografia Quantitativa-2	**2000 a 2005**	Modelos espaço-tempo	Funções	Modelos multi-escala
Geografia Crítica	Segunda década do século 21 (?)	Objetos e Ações Espaço de fluxos e espaço de lugares	*Ontologias e Espaços não-cartográficos*	Representação do Conhecimento

- **Geografia Crítica** (SIGs do futuro): aqui, os conceitos-chaves incluem o espaço como *sistemas de objetos* e *sistemas de ações*. Pode-se apenas especular sobre as representações computacionais que serão utilizadas nesse contexto, que possivelmente incluirão técnicas de representação de conhecimento.

A sistematização da trilogia de CÂMARA (2002) cobriu as obras de autores representativos de diferentes correntes da Geografia. No caso da *Geografia Idiográfica*, cita-se R. Hartshorne. Para a *Geografia Quantitativa* (no Brasil, também chamada de Teorética), citam-se D. Harvey e R. J. Chorley. No caso da *Geografia Crítica*, vêm os trabalhos de Mílton Santos e de D. Harvey.

Por trás de muitas das idéias e conceitos dos geógrafos citados, há contribuições indubitáveis de obras de pensadores e filósofos da atualidade, que procuram abrir caminho em novas áreas de reflexão sobre a natureza da inteligência e das tecnologias da informação, sendo dignos de nota: 1) LÉVY (1990), adaptando as idéias dos cínicos e dos estóicos gregos (*cosmopolitismo*) ao que chamou de *ecologia cognitiva*, fundando uma nova forma de pensar sobre o homem e as tecnologias da informação que o cercam, confundindo-os num

[22] Adaptada de CÂMARA (2002).

dinâmico *coletivo pensante*; 2) MASUDA (1982), numa linha de pensamento parecida com a de LÉVY (1990), analisa os inter-relacionamentos dos grupos sociais que compõem o que chama de *sociedade da informação*; e 3) INGWERSEN (1996), indo além das reflexões, invade os domínios dos sistemas de *recuperação da informação* (SRIs), ao agregar duas dimensões ao projeto dos tradicionais SRIs: a cognitiva e a emocional, demonstrando que somente nesse nível de processamento da informação é possível produzir conhecimento.

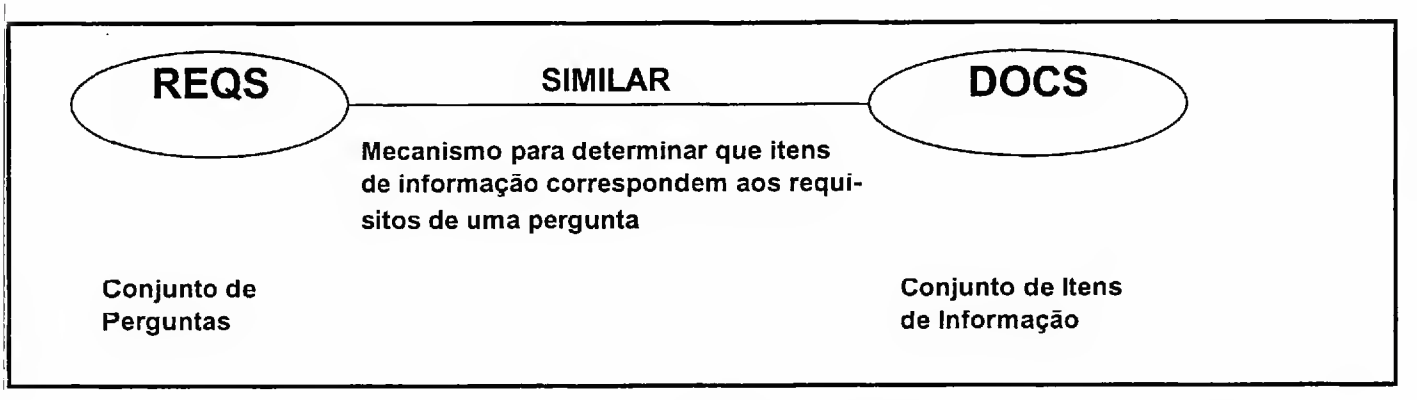

Figura 1.1: Esquema de um sistema de recuperação da informação[23]

INGWERSEN (1996) trata da importância do contexto na fixação de significados de termos relevantes numa consulta e da atribuição de pesos às equações de similaridade, a fim de diminuir as incertezas (ambigüidades) no SRI, entre outras considerações que acabam por autorizá-lo a asseverar, como fez RICHARDSON (1999), que as pesquisas na área da IA, especialmente as tendências dos SIs que exploram aspectos cognitivos do comportamento inteligente, tendem a minimizar as profundas divergências ideológicas que existem entre os métodos quantitativos e qualitativos de pesquisa.

Um assunto que é comum tanto para a Ciência da Informação como para a CIGeo, intimamente ligado ao objeto de ambas, é a *recuperação da informação*.

Como o presente trabalho de pesquisa está muito ligado às teorias e às aplicações sobre recuperação automatizada de informações e documentos[24] e como esta área desenvolveu-se de uma forma intrinsecamente ligada à IA[25], vale a pena fazer uma sinopse, baseando-se no estado-da-arte levado a termo por dois autores: ROBERTSON (1994)[26] e MEDEIROS (1999).

---

[23] Retirada de MEDEIROS (1999).
[24] Mapas, cartas e imagens de satélite também são documentos (GUINCHAT, 1994).
[25] Particularmente na linha de pesquisa chamada de Algoritmos Genéticos (SANTOS, 2001) e PLN.
[26] Enumerou modelos **matemáticos**, **probabilísticos** e **cognitivos** de recuperação automática de informação.

Grosso modo, a recuperação da informação diz respeito à representação, ao armazenamento e à organização de informações, para permitir que os usuários tenham acesso a itens de informação específicos. Constitui-se num conjunto de procedimentos que possibilitam comparar as perguntas feitas pelos usuários com as descrições dos itens armazenados. Com base nessa comparação, determinam-se quais itens são apropriados para determinada solicitação [G. Salton, *apud* MEDEIROS (1999)].

Os sistemas de recuperação da informação podem ser descritos como um conjunto de itens de informação (DOCS), um conjunto de perguntas (REQS) e algum mecanismo (SIMILAR) para determinar que itens de informação correspondem às exigências da pergunta, conforme representado na Figura 1.1.

A recuperação da informação tem suas origens na análise de conteúdo (análise de assunto, análise temática), que remontam à Idade Média, em que o rigor científico não estava presente. Somente em 1640, registrou-se um trabalho feito na Suécia para estudar a autenticidade de noventa hinos religiosos e que já possuía algum valor científico; mas é somente no início do séc. XX, na área de jornalismo, que surgiram as primeiras técnicas realmente de cunho científico para estudos de análise de conteúdo, nos EUA (RICHARDSON, 1999).

A análise de conteúdo é muito utilizada no âmbito das ciências sociais aplicadas para analisar, descrever e classificar textos ou discursos. Na Ciência da Informação, o termo ganha outra dimensão (mais abrangente) pelo foco nos *documentos*, em substituição às tradicionais unidades de estudo como textos e discursos. Portanto, é lógico que esse campo de estudo ganhasse outra denominação no âmbito da CI, qual seja: *Análise Documentária*, ao tratar de documentos e da utilização de análises, descrições e classificação para recuperar o conteúdo desses documentos. Segundo M. L. G. Lara [*apud* MEDEIROS (1999)], a introdução do termo *análise documentária* (AD) na literatura da área de informação deve-se, em grande parte, a Jean-Claude Gardin.

MEDEIROS (1999) registrou que, com base na análise das revisões de literatura publicadas no *Annual Review of Information Science* (1969 a 1989), concluiu-se que a questão da AD é objeto de esforço contínuo nas pesquisas da área de Ciência da Informação. A autora salientou duas considerações de autores nesse assunto, que resumem a evolução das técnicas tradicionais de AD para as de PLN (processamento de linguagem natural). As primeiras considerações são de F. W. Lancaster:

"Por mais de trinta anos, pesquisadores procuram maneiras de substituir o processamento intelectual do homem por processamento de textos por computador. Os métodos in-

vestigados incluem indexação automática por extração ou atribuição; extração automática; organização automática de termos ou documentos em classes; automação de construção de tesauros ou desenvolvimento de redes semânticas de relações entre termos; e métodos de comparar perguntas em linguagem natural a textos de documentos ou suas representações."

As próximas considerações são de D. M. Liston e M. L. Howder, que justificam essa natureza heterogênea da área:

"...isso demonstra o quanto o campo é complexo, pelo fato de estarmos lidando com algo relacionado à compreensão de como pensamos, como formulamos conceitos, como os comunicamos, como os registramos, como os salvamos para uso futuro, como os utilizamos, como os reutilizamos em permutações e combinações ilimitadas para sintetizar ou elaborar novos conceitos."

Conclui-se que, em sua essência e natureza, o processo permanece o mesmo: há necessidade de análise da informação registrada e de extração de seu conteúdo de forma a possibilitar sua recuperação. No entanto, modificam-se amplamente as ferramentas e técnicas utilizadas no processo, sobretudo com a introdução de novas tecnologias.

A Ciência da Informação não dispõe de toda base teórica e metodológica para resolver os problemas da análise documentária, por isso busca, notadamente nas áreas de Lingüística, Ciência da Computação, Matemática e Probabilidade e Estatística, referencial teórico e técnico para o desenvolvimento de procedimentos de tratamento de conteúdo.

Não se pretende, neste trabalho, analisar o conjunto de métodos e técnicas de AD apresentados na literatura, mas dentro do espírito de construir um cenário histórico, que justifique a desembocadura das várias técnicas e teorias tributárias para esse domínio interdisciplinar da similaridade semântica (SS), cabe assinalar extensões da AD que têm sido muito discutidas em meios de comunicação científica modernos. Essas técnicas são: a Análise Fatorial[27] (AF) e a Análise de Componentes Principais (PCA[28]). Tratam-se de fórmulas numéricas, utilizadas para reduzir a complexidade de problemas das áreas de pesquisa qualitativa, diminuindo drasticamente o conjunto original de variáveis estabelecidas normalmente de forma arbitrária ou subjetiva.

A recuperação da informação por AD, vista pela IA, recebe a atenção de um dos mais complexos porém mais explorados de seus campos de estudo: o processamento automático de linguagem natural (PLN[29]).

---

[27] Do inglês: *Factor Analysis*

[28] Para não dissociar o uso consagrado da sigla desta técnica, até em trabalhos no vernáculo, manter-se-á a original, em inglês, que significa: *Principal Component Analysis*.

[29] Tratamento Automático de Linguagem Natural (TLN), conforme MEDEIROS (1999).

As primeiras idéias sobre PLN remontam à década de 30, com o surgimento de pesquisas voltadas para a tradução automática (KONDRATOV, 1972).

Apesar do fracasso experimentado nas pesquisas iniciais, a década de 50 foi marcada pelo surgimento de idéias fundamentais para o desenvolvimento futuro de sistemas de PLN, como atestaram os trabalhos de J. L. Harris e Noam Chomsky, relativos à gramática formal e à gramática *transformacional*, respectivamente, e pelo surgimento da IA.

Na década de 60, motivados pelos resultados desses trabalhos, foram criados sistemas de PLN que alcançaram resultados representativos para a época. Esses sistemas baseavam-se, sobretudo, nos aspectos formais da linguagem, sem relacioná-los ao significado. A dificuldade dos sistemas para tratar aspectos da linguagem que necessitam de informação semântica só ocorreu de forma efetiva na década de 70.

Os sistemas de PLN evoluíram, nos últimos anos, segundo o nível de tratamento lingüístico, de sistemas que aplicavam conhecimento morfológico e sintático, a sistemas que utilizam conhecimentos sintáticos, semânticos e pragmáticos, no âmbito das ciências cognitivas.

E para encerrar este item, vale a pena tecer mais considerações sobre um assunto já tratado superficialmente na introdução: ***ontologias***.

Novos conceitos (*constructos*) surgiram muito recentemente (ainda não param de surgir!) na Ciência da Computação (IA), quando foi preciso interligar dois ou mais programas de estruturas bem distintas e manter um elo de comunicação com os seus usuários (interfaces gráficas). A necessidade de desenvolver um vocabulário comum para as aplicações sobre as quais esses programas atuavam foi o ponto de partida para a criação de *ontologias*. Daí, evoluindo de coleções de termos bem controlados, elas passaram a ter o objetivo de descrever as visões de mundo de certos grupos de usuários e de armazenar o conhecimento desses grupos.

A moderna sociedade da informação tem nessas pesquisas sobre interfaces gráficas e nas camadas de representação do conhecimento por trás delas (as ontologias), um meio seguro para consolidar-se frente à enxurrada de informações produzidas todo o dia.

Numa fase preliminar da metodologia, não se formalizou uma ontologia, porque se trabalhou com um protótipo (PROFAX) baseado em algoritmos da programação estruturada e com alguns elementos de OO, para fornecer entendimento mínimo sobre a SS. No entanto, as declarações em determinada LTP (*Java*™, no caso) para esse protótipo acabaram incorporando algumas características de uma LRC (linguagens de representação do conhecimento) para o modelo que se deseja conceber, mais adaptadas para construir ontologias.

Mesmo fugindo dos requisitos (ainda um tanto nebulosos) de elaboração de uma ontologia, pode-se dizer que um algoritmo transformado num conjunto de declarações de determinada linguagem de programação (LTP) não deixa de ser uma linguagem de representação do conhecimento, ainda quando reduz a complexidade das formulações da Lógica de Primeira Ordem típicas de uma ontologia, cujo fim é representar o conhecimento da área de SS, aos cânones de BÖHM (1966) e DIJKSTRA (1968), baseados em declarações de LTPs, as quais se assentam no trinômio: *seqüência* (*top-down*) – *alternativa* (*if-then-else*) – *repetição* (*while-do* e outras).

1.3. Identificação e enquadramento da pesquisa

O presente trabalho é um estudo-de-caso exploratório, de caráter híbrido (tanto de aporte quantitativo como qualitativo).

Na seção dedicada à apresentação da metodologia (Cap. 6), será descrita em maior minúcia a natureza mais geral do método científico empregado neste estudo-de-caso, que segue, em linhas gerais, a tipologia de TRIPODI (1975): pesquisa exploratória de manipulação experimental.

Esta pesquisa teve origem nas teorias que vêm sendo levantadas pelo grupo de estudos de Ciência da Informação Espacial, liderado pelo Professor Max J. Egenhofer, da Universidade do Maine (EUA). O problema fundamental surgiu da quebra de expectativas desse grupo em relação aos SIGs tradicionais, de bases eminentemente formais (lógico-matemáticas) e materializados por tecnologias da informação nada fáceis de manipular (até para especialistas).

A definição de problemas e a formulação de hipótese no campo da similaridade semântica de classes de objetos espaciais servirá de roteiro básico para a escolha de que observações devem ser estudadas e quais experimentos devem ser realizados. A revisão de literatura gira em torno da formulação (modelo matemático) estabelecida na tese de RODRÍGUEZ (2000), o marco teórico no qual a presente pesquisa se sustenta e prossegue.

O modelo concebido nesta tese, inspirado no MSS (modelo de similaridade semântica) de RODRÍGUEZ (2000), têm uma função tríplice: 1) Coletor de dados (termos) de um subconjunto do modelo conceitual da carta topográfica da região de Faxinal (PR); 2) Processador desses dados para avaliar a SS das classes de entidades espaciais denotadas por esses termos, e 3) Base de corroboração de hipóteses, juntamente com os instrumentos de sondagem (questionários) aplicados a profissionais do *Geoprocessamento*.

Já que se mencionou o termo "protótipo", é bom que fique clara a sua acepção neste estudo e isto vai ser levado a efeito no subitem 6.3.1 (metodologia). Vale a pena, por enquanto, ficar com a finalidade tríplice, descrita *ut retro*, e declarar que um protótipo não é um sistema; *grosso modo*, é um "quase-sistema".

A expectativa para esta pesquisa é que esse protótipo seja um produto adequado para confirmar as hipóteses (pressupostos) levantadas e, ao mesmo tempo, seja viável diante das limitações de ordem logística e cronológica, normais de uma pesquisa de doutorado, sem preocupações excessivas com o rigor de análise e projeto de sistemas. Esses cuidados poderão ficar para trabalhos derivados deste, que serão propostos no Capítulo 7.

Sintetizando, o objetivo desse protótipo é conferir mais resistência à conjectura de que é possível criar um programa de computador que simule a capacidade humana para reconhecer semelhanças (ou diferenças) entre fenômenos da superfície terrestre (realidade física), modeláveis segundo alguma técnica (UML™, p.ex.).

1.3.1. Área de concentração: Transferência da Informação.

Esta é a única área de concentração oferecida pelo Departamento de Ciência da Informação e Documentação (CID) da Universidade de Brasília (UnB) para o curso de doutorado.

A Transferência da Informação do CID/UnB coaduna-se com áreas afins de outros cursos de pós-graduação (*stricto* e *lato sensu*) em Ciências Exatas e da Terra, com os quais a Ciência da Informação (CI) mantém interdisciplinaridade[30]. Nesses cursos de pós-graduação, por exemplo, a área de concentração é Cartografia, que também tem como um dos seus objetivos transferir informação, especialmente na atual conjuntura, em que a interoperabilidade entre SIGs tem sido o problema central de muitas pesquisas.

1.3.2. Linha de pesquisa: Processos e linguagens de indexação.

São quatro as linhas de pesquisa oferecidas pelo CID/UnB: 1) Planejamento, administração, gerência e avaliação de bibliotecas e sistemas de informação; 2) Processos e linguagens de indexação; 3) Formação profissional e mercado de trabalho; e 4) Comunicação científica.

Processos e linguagens de indexação é a linha de pesquisa que trata de: análise da informação e processos de indexação; organização do conhecimento; análise de conteúdos; processos de classificação, indexação e linguagens documentárias; Terminologia; aplicações da Informática e indexação automática.

A segunda linha foi escolhida para enquadrar este trabalho, porque contém mais atributos que se coadunam com o objetivo geral da pesquisa, particularmente no que tange à interoperabilidade entre SIGs. Esta linha de pesquisa mantém a necessária correlação com vários cursos de doutorado em Ciências Exatas e da Terra.

1.4. O tema da pesquisa

Já é bem desconfortável inserir um trabalho de pesquisa num quadro de teorias científicas que se supõe estruturado, ou que não suscite grandes controvérsias entre seus pares cientistas. No caso em tela, não tanto pelas Ciências da Terra (*geociências*), mas pelas chamadas ciências cognitivas, a tarefa de situar o trabalho tornou-se, efetivamente, uma das que mais tempo exigiu nesta pesquisa, perdendo apenas para a construção de um dos instrumentos de sustentação da hipótese de pesquisa - o PRONTO©: PROtótipo de avaliação da similaridade semântica entre classes de entidades espaciais, representadas numa ONTOlogia *ad-hoc*.

De forma bem abrangente, o tema proposto encontra-se na interseção dos campos científicos da Cartografia e da Ciência da Computação, para resolver um problema comum de natureza comunicativa, em sua essência. Portanto, é interdisciplinar.

Descendo mais no nível de especialização desses campos, sem impor limites precisos no problema interdisciplinar que logo será formulado, o tema está na interseção da disciplina de SIG, grande tributária da Cartografia, com a disciplina de Inteligência Artificial (IA), grande tributária da Ciência da Computação.

A dificuldade mencionada anteriormente está em se analisar o quadro em que se insere o assunto na Ciência da Computação e onde se insere a própria Ciência da Computação no contexto atual de ciência pós-moderna, particularmente por causa de uma de suas disciplinas: a IA. No caso da Modelagem de Objetos[31], também disciplina da Ciência da Computação e parte do problema de pesquisa que se formulará a seguir, não há dificuldade similar à apresentada pela IA.

Não faz parte do presente estudo encetar uma incursão profunda (e palpitante) nos vários tratados e reflexões, que não param de surgir, na tentativa de institucionalizar uma nova área multidisciplinar voltada para o estudo científico da mente: a *Ciência Cognitiva*. Acompanhando a cautela de ABRANTES (1994), o texto desta tese utiliza o termo plural e de le-

---

[30] V. glossário.

[31] Padronizada internacionalmente pela UML™ (*Unified Modeling Language*).

tras iniciais minúsculas – ciências cognitivas -, que abrange contribuições da Ciência da Computação, da Lingüística, da Psicologia, da Neurofisiologia e da Epistemologia. A explicação (e não uma justificativa) para empregar "ciências cognitivas", assim como outras considerações de ordem epistemológica, serão apresentadas no subitem 1.5.4 (As inteligências).

Mesmo não sendo o objetivo desta tese discutir a autonomia e a unidade metodológica desta área do conhecimento científico que emerge, é da inteira responsabilidade do pesquisador o espaço que foi dedicado no subitem 1.5.4, a fim de pôr o leitor interessado a par do que se bate por trás do sumo em revisão de literatura extraído para este texto. É uma seção da tese que pode ser desconsiderada por leitores mais familiarizados com o assunto e que já tenham formado alguma opinião sobre qual tendência aderir. Para estes, é importante que já tenham a posição deste pesquisador: a *linha cognitivista* da Inteligência Artificial (IA), que vem da *corrente funcionalista* da Filosofia da Mente.

Sem desejar uma simplificação excessiva do complexo assunto sobre a natureza da inteligência neste curto subitem, a linha de pensamento em que se desenvolve esta tese é uma entre três, categorizadas como enfoques sobre a natureza da inteligência. Todos os enfoques têm pontos fortes e limitações; todos os seus defensores criticam e recebem críticas, mas nenhum deles é tão polêmico como o *cognitivista.*

O subitem 1.5.4 tratará com mais minúcia deste e dos outros dois enfoques. Essa classificação triádica vem de Roger Schank e Lawrence Birbaum [*apud* Khalfa (1995)], que classificaram a linha de pensamento cognitivo-funcionalista como *"posição de inteligência aditiva"*.

O outro enfoque é o de Noam Chomsky e outros lingüistas, chamado de *"posição do órgão de linguagem"*, que sustenta a exclusividade das capacidades intelectuais para o ser humano, dotado de órgãos específicos para a fonação e, portanto, capaz de se expressar e de se comunicar com outros seres humanos por meio de um sistema lingüístico.

Finalmente, o terceiro enfoque é o do filósofo John Searle, chamado de *"posição da consciência do cão"*, no qual ele argumenta que só pode haver indício de inteligência em seres orgânicos e dotados de consciência. Um cão poderia ser inteligente, mas não uma máquina.

A tese cognitivo-funcionalista funda-se nos chamados "estados mentais" (instâncias), caracterizados por inter-relações funcionais de um cérebro, não necessariamente humano. É sobre esta abstração que recaem as maiores críticas de outras duas origens (além dos defensores dos outros dois enfoques): de forma mais veemente, vêm as críticas das correntes de fundo *behaviorista* (*comportamentalista*) da Psicologia, que sustentam serem inacessí-

veis aos métodos científicos conceitos de "estados mentais", "crenças", "juízos de valor" e "representações mentais", se não for possível explicá-los em termos de funções cerebrais observáveis e verificáveis; de outro lado, mais transigentes, dentro do próprio campo da IA, vêm os *conexionistas*, questionando a instanciação dos "estados mentais" em estruturas não-biológicas, como defendem os cognitivistas. Para os conexionistas é difícil admitir que estruturas de circuitos integrados de elementos semicondutores como o silício (Si) ou o germânio (Ge) possam replicar processos mentais humanos.

Vale ressaltar que as noções conexionistas formam uma espécie de meio-termo entre as teorias cognitivas da IA e as teorias comportamentalistas. Não há teorias conexionistas consagradas, porque o domínio de problemas em que se aplicam é muito restrito, no entanto, dos resultados práticos obtidos com suas redes neurais artificiais (RNA), capazes de realizar aprendizado associativo e de simular em mecanismos físicos as conexões sinápticas do sistema nervoso de algumas espécies do reino animal, incluindo o homem, é bem possível que, em breve, essa ponte que o conexionismo faz entre a sua coirmã cognitivista e as validadas teorias neurofisiológicas e psicológicas venha a produzir ricos e fecundos subcampos de estudo na IA, corrijam seus rumos e até estendam o alcance desses benefícios a outros campos do conhecimento que, com a IA, compartilhem problemas .

No caso da *informação geográfica*, é necessário enfocar esse "berço" de problemas para os quais a tecnologia sozinha não conseguiu oferecer soluções adequadas, pelo contrário, produziu sistemas de tratamento dessa informação que não permitem um intercâmbio de dados coerente e confiável.

É especialmente sobre os aspectos de ordem semântica da informação geográfica (BÄHR, 1996) que surgiu o tema. Esses aspectos constituem fatores que afetam direta ou indiretamente a representação da informação geográfica. São fenômenos que podem ser observados por enfoques de ordem lingüística (foco no signo verbal) e mesmo de ordem semiótica (foco no signo em sua expressão mais completa: iconográfica, indicativa e simbólica). Vê-se, por conseguinte, que os problemas de informação geográfica acabam imbricados nos das ciências cognitivas.

Durante a fase que antecede a revisão de literatura, no levantamento bibliográfico, de características muito mais abrangentes, já se verificava que tanto os pesquisadores da área de CI quanto os da área de SIG e de Ciência da Computação tinham e continuam tendo preocupações semelhantes na teorização e no tratamento da informação geográfica, ao procurem um meio de torná-la acessível e simples de ser manuseada por um usuário que não seja um especialista nas tecnologias dessas áreas. Por conseguinte, essa convergência de interesses e de questionamentos que os artigos coligidos dessas três áreas indicavam

interesses e de questionamentos que os artigos coligidos dessas três áreas indicavam constituiu mais uma causa do interesse por essa pesquisa.

O interesse pelo tema também despertou da leitura do trabalho desenvolvido pela pesquisadora chilena RODRÍGUEZ (2000), da Universidade do Maine (EUA), que em muitos aspectos aprofundou algumas das reflexões de BÄHR (1996) e que respondeu a muitas das questões por ele levantadas.

O trabalho de RODRÍGUEZ (2000) é o marco teórico desta tese. Nesse trabalho e em diversos contactos com a sua autora, ficou sempre patente o enquadramento do tema desta tese na área de concentração de recuperação (descoberta) da informação. A declaração seguinte da autora sintetiza o referencial teórico em que laborou:

"Hoje, as pessoas e organizações não estão muito interessadas em localizar dados; elas querem é descobrir informação." (RODRÍGUEZ, 2000)

Entretanto, o acervo de dados acumulados nessas últimas décadas é descomunal e não pode ser descartado.

Vive-se nos bastidores da incompatibilidade entre sistemas de informação geográfica. Essa incompatibilidade é a causa da dificuldade de identificação de objetos semanticamente semelhantes em bases de dados heterogêneas e que os tradicionais SIGs não conseguem resolver, apesar do forte embasamento formal[32] que possuem e das infindáveis camadas de programas, às vezes não muito amigáveis (*user-friendly*), que se lhes acrescem.

O tema, dessa forma, vem expresso pelo título do trabalho, constituindo-se num estudo-de-caso sobre a avaliação da similaridade semântica entre classes de entidades espaciais que "povoam" (ou que podem "povoar") uma base de dados cartográficos (ou espaciais, em sentido lato). Esses dados são de uma região do sul do Brasil (Faxinal – PR) e foram obtidos das mais diversas formas (levantamentos topográficos, tomadas de fotos aéreas, recobrimento por imagens de satélite, etc.), seguindo todo o rigor previsto nas normas técnicas vigentes.

As técnicas de tratamento de gabinete foram efetuadas pelos mais avançados recursos de *hardware* e *software* de *geoprocessamento* corporativos, sob a responsabilidade da 1ª Divisão de Levantamento (1ª DL), subordinada à Diretoria de Serviço Geográfico do Exército Brasileiro (DSG). A estruturação desses dados seguiu o paradigma da orientação a objeto (OO), em fase de desenvolvimento na DSG, a fim de preparar o seu parque de produção de documentos cartográficos para as novas exigências de precisão, economia e rapidez no a-

[32] Lógico-matemático.

tendimento das necessidades de usuários de perfis cada vez mais variados e dependentes de mapas, cartas e plantas em base celulósica (papel) ou digital.

1.5. Definições preliminares

Como são poucos os trabalhos até agora apresentados no curso de pós-graduação em CI da Universidade de Brasília (UnB), que tenham por objeto de estudo a informação geográfica, é imprescindível cercar a apresentação do problema com esta ambientação aos problemas e objeto de estudo da Cartografia e dos SIGs, assim como da Ciência da Computação, especificamente da IA.

Nesse ponto, o que mais importa explicar para tornar o trabalho coerente com o seu propósito subjacente (descobrir, recuperar informação) é antecipar algumas definições de termos que surgirão, logo a seguir, nas etapas de formulação do problema, nos enunciados das hipóteses e no estabelecimento dos objetivos, evitando lapsos de entendimento para o leitor ou controlando a indesejável ambigüidade[33].

Outros aspectos dessas definições e mais outras que as circundam serão examinados noutras seções do trabalho. Aqui, o objetivo é de pôr o leitor em condições de iniciar exercícios de denotação, i.e., de relacionar os termos dessas línguas profissionais com as referências do mundo-real geográfico (referência empírica).

Em razão da exaustiva utilização de alguns termos que, a rigor, expressam-se por definições distintas, fica este registro preliminar, que será relembrado diversas vezes ao longo do texto, com as necessárias explanações. Esses pares de termos de discriminação relaxada são os seguintes: *entidade* ou *objeto*, *cartográfico* ou *geográfico*, *banco de dados orientado a objetos* ou *banco de objetos*.

1.5.1. Informação geográfica (IG)

Trabalhar com IG significa, antes de mais nada, *utilizar computadores* como instrumentos de representação de dados atrelados a um certo sistema de referência espacial. Desse modo, o problema fundamental da CIGeo é o estudo e a implementação de diferentes formas de representação computacional do espaço geográfico (CÂMARA, 2002).

É costume dizer-se que o *Geoprocessamento* tem natureza interdisciplinar, ou ainda, que *"o espaço é uma linguagem comum"* para as diferentes disciplinas do conhecimento. Apesar de válidas num sentido genérico, essas noções escondem uma minúcia que necessi-

ta de esclarecimento: a pretensa interdisciplinaridade dos SIGs é obtida pela redução dos conceitos de cada disciplina a algoritmos e estruturas de dados, utilizados para armazenamento e tratamento dos dados geográficos. Essa nuança ficará clara com a citação de alguns problemas típicos de vários campos do conhecimento e das atividades humanas:

- Um *sociólogo* deseja utilizar um SIG para entender e quantificar o fenômeno *da exclusão social* numa grande cidade brasileira.
- Um *ecólogo* usa um SIG com o objetivo de compreender as *florestas remanescentes* da Mata Atlântica.
- Um *geólogo* pretende usar um SIG para determinar a *distribuição de um mineral* numa área de prospecção, ao valer-se de um conjunto de amostras de campo.
- Um *chefe militar* usa um SIG, no seu *carro de combate*, ao visualizar uma carta eletrônica num monitor de vídeo e ao obter informações sobre áreas de inundação do terreno que estão em seu rumo, evitando atolar-se.

O que há de comum em todos os casos acima? Para começar, cada especialista lida com conceitos de sua disciplina (exclusão social, desflorestamento, distribuição mineral, transitabilidade de veículos blindados).

Para utilizar um SIG é preciso que cada especialista transforme conceitos de sua disciplina em representações computacionais. Após essa tradução, torna-se viável compartilhar os dados de estudo com outros especialistas (eventualmente de disciplinas diferentes). Noutras palavras, quando se diz que o espaço é uma linguagem comum no uso de SIG, faz-se referência ao espaço representável em programas de computador e não aos conceitos abstratos de espaço geográfico.

Tais cenários de aplicação situam esta pesquisa no seu domínio temático, manancial de diversos problemas ligados à representação computacional de dados geográficos, quais sejam: *Como representar, em computadores, os dados geográficos? Como as estruturas de dados geométricas e alfanuméricas se relacionam com os dados do mundo-real? Que alternativas de representação computacional existem para dados geográficos?*

Diante da gama extensa de problemas que são colocados para a CIGeo, a IG não poderia cingir-se tão-somente ao exame da informação contida nos signos verbais (unidimensionais[34]). Ela é uma informação de caráter dual, que associa texto a feições gráficas (multidimensionais), sendo obtida do cruzamento de dados quantitativos e qualitativos colhidos de

---

[33] De Arthur Bachrach: "A suposição de mútua compreensão é uma visão imatura do método científico. Dizer que todos sabem é repetir a pergunta e evitar o assunto principal da clareza e precisão científicas da definição".

[34] Segundo SAUSSURE (1975), é a dimensão linear do significante auditivo, que se manifesta cadenciado ao longo do tempo ou formando uma cadeia linear e seqüencial, quando expresso em linhas escritas num papel.

campo ou interpretados, como por exemplo: de medidas de ângulos, distâncias, assim como de fotografias aéreas e terrestres, toponímia e imagens de *sensores orbitais* ativos e passivos e a interpretação dessas fotografias e imagens (aspecto qualitativo).

Os *sensores orbitais* (satélites artificiais) ativos, citados anteriormente, são aqueles que detectam um objeto espacial da superfície terrestre pela emissão de um feixe de ondas (microondas) de sua própria fonte de energia – radares ou sensores ativos. Os passivos, ao contrário dos ativos, não possuem fonte própria de emissão de energia para iluminar um objeto afastado e localizado na superfície terrestre. Para detectá-lo, o sensor utiliza uma fonte de iluminação externa - o sol, por exemplo. (V. *sensores remotos* no glossário).

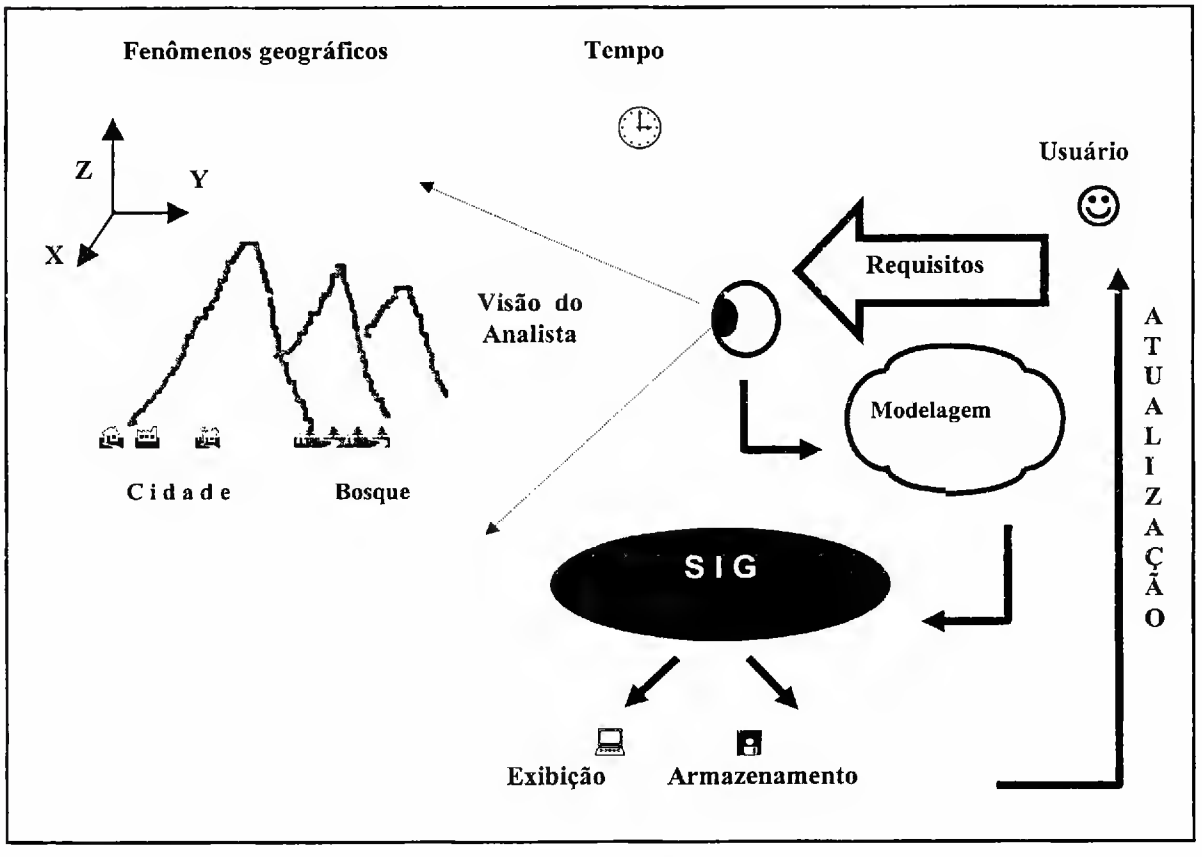

Figura 1.2: Como surge e é modelada a informação geográfica.

A *toponímia* é o componente mais trivial (pelo enfoque clássico de SIs) da IG: o *texto*. Ela é um conjunto de termos designativos de acidentes naturais e artificiais encontrados na superfície terrestre, atribuídos por equipes especializadas em *reambulação*[35] (do latim *re* – 'repetição, intensidade' – e *ambulare* – 'caminhar'). É sobre essa toponímia que recairá a

maior parte do esforço de verificação exploratório-experimental desta tese, uma vez que o protótipo que se pretende construir é baseado em relações conceituais entre esses termos, cujas técnicas de exploração são muito similares entre a OO e a Análise Documentária da Biblioteconomia, segundo SILVA (2001).

O componente menos trivial da IG – o *espacial* - é o que se refere à transformação das medidas numéricas de ângulos e distâncias geométricas das (ou entre as) entidades espaciais da superfície terrestre. Por meio de entes abstratos chamados de vetores, formalizáveis pela Matemática e representáveis por algoritmos computacionais, esse componente está sempre atrelado a um sistema de posição ou de referência, num espaço *n-dimensional*. Normalmente, este sistema é tridimensional, como um sistema de coordenadas retangulares, representado pelo terno ordenado [X,Y,Z]. Não é do interesse desta pesquisa recuperar esse componente da base de dados que se examinará, mas é preciso descrevê-lo, porque este trabalho deve ser prosseguido de forma a prever a captura de componentes espaciais representados geometricamente.

A Figura 1.2 mostra o fenômeno geográfico sendo observado por um analista humano, não só considerando as variáveis espaciais mas também o tempo, que condiciona o grau de atualização da informação geográfica. Além do obsoletismo que afeta a IG, a ambigüidade é outro fator que deve ser considerado no desenvolvimento de um moderno SIG.

Atualmente, o modo de ver a informação geográfica é menos pela imposição da abstração geométrica aos fenômenos físicos do mundo-real e mais pela admissão de técnicas avançadas de modelagem desses fenômenos, como a orientada a objetos, por exemplo. Essas técnicas, implementadas por LTPs estruturadas, são mais adequadas aos métodos cognitivos de percepção do mundo-real (MR) e para representar outros tipos de relações (generalização e agregação, p.ex.) entre as entidades espaciais, que diferem dos tradicionais princípios da Geometria Euclidiana.

Foi mencionado o termo *ambigüidade*, num contexto de aplicação espacial. Apesar dos efeitos indesejáveis na representação final de documentos cartográficos que o significado desse termo representa, sua importância neste trabalho é secundária. Na breve inspeção sobre os efeitos da ambigüidade sobre as fases por que passam os dados cartográficos até se transformarem em IG, mas ficará claro que é por causa dos efeitos dessa ambigüidade que a meta de interoperabilidade entre sistemas de informação geográfica ainda não foi alcançada.

---

[35] V. glossário.

PRADO (2000), citando trechos dos trabalhos de PRATT (2000), Raul Ramirez, BERTIN (1967) e CHOMSKY (1998), trabalha com a idéia de existir muito em comum entre *mapas* e outras formas de *comunicação*, como línguas naturais (na forma escrita). Ambas as formas de expressão utilizam marcas de tinta sobre o papel, por intermédio de um código preestabelecido. Entretanto, o que diferencia os mapas é a sua característica geométrica e visual, aproximando-os dos objetos que se deseja representar.

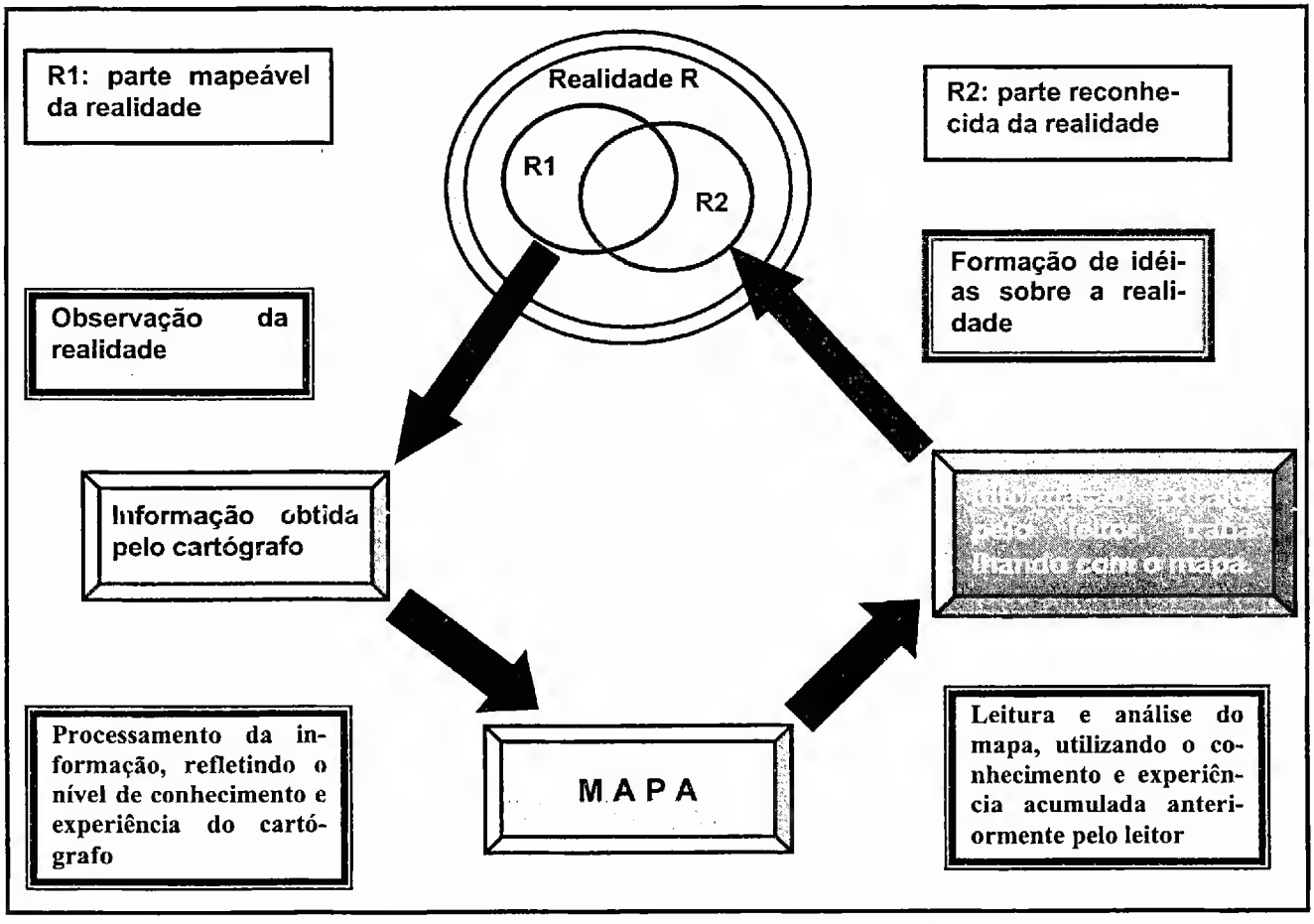

Figura 1.3: A comunicação cartográfica.[36]

Para aclarar mais essa analogia entre a *ambigüidade* que afeta o signo verbal e o não-verbal, é instrutivo apreciar a definição dada ao termo no trabalho de MEDEIROS (1999), que explorou as possibilidades de sistemas especialistas no tratamento de linguagens naturais, envolvendo conceitos da Lingüística e da IA. A pesquisadora explicou que entre os fenômenos estudados na Lingüística que devem ser solucionados por sistemas especialistas baseados em IA encontra-se o da ambigüidade, que interfere no tratamento do conteúdo dos documentos e na recuperação da informação.

[36] Adaptada de MARTINELLI (1991).

No caso específico da recuperação de informação textual, esse efeito indesejável traduz-se pela inexistência de correspondência exata entre o conjunto de palavras ou frases e o conjunto de significados, quando se consideram os sistemas automatizados de recuperação da informação em linguagem natural. Para esse processo de recuperação, objetivo precípuo da análise da linguagem natural, não importa apenas a extração de palavras do texto, mas fundamentalmente as relações de significado que estas palavras mantêm no contexto do discurso em que ocorrem. Sistemas de recuperação (p.ex., de indexação automática) que não se utilizam de conhecimentos lingüísticos, como os que extraem palavras dos textos e os que se baseiam em métodos estocásticos, não podem fazer referência a um objeto da realidade extralingüística do autor do documento. Não podem exprimir o que "dizem" os documentos, porque tratam as palavras do texto isoladamente, desconsiderando o contexto em que ocorrem [adaptado de R. Bouché, citado em MEDEIROS (1999)].

No caso específico da recuperação de informação geográfica, há vários níveis de abstração, que também são explicados pela Teoria Semiótica de Charles Sanders Peirce. Esses níveis vão desde a impressão (contemplação) do observador diante de um fato de natureza geográfica (ainda não é um fenômeno), passando pela percepção desse fato (aí já é um fenômeno) e chegando até a representação mental ou material[37] desse fenômeno (interpretação), que pode ser um processo infinitamente recorrente, isto é, uma interpretação gera uma outra e assim sucessivamente (SANTAELLA, 1983). É a gênese de um sistema de conceitos pelo enfoque da Semiótica, que se ajusta adequadamente ao processo de comunicação cartográfica.

A idéia de linguagem cartográfica surgiu de vários autores (V. subitem 1.2.2), para os quais os mapas são entendidos como veículos de comunicação (*comunicação cartográfica*). Fazer um mapa significa desencadear este processo de comunicação (V. subitem 3.2.2.2), sempre tendo em vista que, se há comunicação, há ruído (ambigüidade).

No âmbito do Serviço Geográfico do Exército Brasileiro (SGE), como noutros órgãos congêneres de outros países, de tradição até secular[38], não é inusitado o entendimento de que um mapa ou uma carta topográfica é uma modalidade de linguagem, não importando o meio (celulósico ou digital) em que se dissemina a informação. Nesse aspecto, é elucidativo citar o preâmbulo de um manual técnico proposto pela DSG, que trata justamente das regras de ordem sintática, semântica e pragmática para esse tipo de linguagem:

---

[37] Por um símbolo ou convenção cartográfica.

[38] O brasileiro opera nesse código lingüístico desde 1890. O português, desde o séc. XV, na Escola de Sagres (RAISZ, 1969, p.19-35)..

"a. A carta topográfica é a representação gráfica e simbólica do terreno. Deve refletir fielmente o aspecto físico da área levantada e as obras humanas que o terreno possibilitou ou condicionou.

b. A nomenclatura geográfica, composta de topônimos e antropônimos, é uma das partes mais importantes e delicadas da carta, porque, aí lançados, animam e, em síntese, registram a linguagem essencial falada na região representada. No âmbito cartográfico, a toponímia é como um registro civil da região. Com efeito, eliminem-se da carta os topônimos e a representação da área e ela se torna inerte e incógnita, apesar de todo seu enquadramento analítico.

c. O estudo toponímico deve explorar o embasamento constituído pelas camadas lingüísticas estratificadas. No Brasil, é necessário o estudo da camada pré-cabralina, sobretudo do primitivo tronco tupi-guarani, e das camadas ameríndias pós-cabralinas, sobretudo do tupi e do guarani, bem como do português e dos derivados das línguas africanas e européias." (BRASIL, 1998a, p. 1-1)

Esse processo de comunicação (V. Figura 1.3) realiza-se em etapas. Ele reúne atividades de elaboração e de uso. O seu entendimento garante ao cartógrafo uma conscientização do papel social de sua profissão. Não existe a neutralidade do construtor de mapa. Ele é um cidadão que presta um serviço à coletividade, que precisa estar próximo das pessoas e do meio em que vivem. Tanto as pessoas como o meio em que vivem são objeto das representações cartográficas e também cabe ao cartógrafo conscientizar tais pessoas sobre o que esse veículo de comunicação – o mapa – poderá fazer de útil por elas.

O processo de comunicação cartográfica não deve ficar preso exclusivamente à preocupação da Teoria da Informação em minimizar as perdas quantitativas de dados em cada etapa do processo informativo. Daí surgiu um novo campo de estudos: a *Semiologia Gráfica*[39], introduzida por Jacques Bertin, em 1967, com duas subdivisões bem independentes de aplicação: o *tratamento* e a *comunicação*. O surgimento desse campo do conhecimento (vinculado à Lingüística) já havia sido previsto pelo lingüista genebrino Ferdinand de Saussure, desde o início do século passado.

Em sua essência, uma representação gráfica é um sistema semiológico monossêmico[40], quando se está na fase da percepção do fato (concreto ou não), fora do *grafismo*. No campo das percepções das relações entre os sinais, não existe convenção. As variações

---

[39] A Profª. Cecília Westphalen, da UFPR, traduziu para *Neográfica* o termo *Semiologie Graphique* de BERTIN (1967).
[40] Na Lingüística: "Para um termo, um só significado".

visuais estão livres de ambigüidade. É aí que reside o interesse da *Neográfica*, com os seus três princípios fundamentais: a *similaridade*, a *ordem* e a *proporcionalidade*.

É preciso substituir a noção de quantidade de informações pela noção de níveis de informação, que se exprimem verbalmente e graficamente. Informação pressupõe relação. Toda construção gráfica que não permite encontrar as relações essenciais entre os dados pode estar afetada de ambigüidade.

Na utilização dos mapas, o usuário é estimulado intelectualmente não apenas no processo de percepção imediata de estímulos visuais; o processo envolve também a motivação, a atenção, a reflexão e a memória [(BERTIN, 1967) e (MARTINELLI, 1991)]. Assim, a Cartografia tem um forte componente cognitivo, até hoje não muito bem explorado.

Seja o seguinte exemplo de ambigüidade na representação gráfica: Como representar uma escola num mapa? por um pequeno círculo negro, encimado por uma bandeirola triangular ou só pela bandeirola? Quem sabe qual é o melhor símbolo? Existe uma infinidade de "boas" representações. Isso não compete à *Neográfica*. É puro *grafismo*. Agora, se uma escola acolhe o dobro de alunos que uma outra, sua vizinha, nesse mapa, e se esta relação quantitativa de alunos é relevante para a construção de um tema, então já se está diante de um problema *neográfico*, que estabeleceria relações de similaridade, ordem e proporcionalidade para satisfazer a esta relação de maneira gráfica, no mapa. A ambigüidade, portanto, é parte inerente ao *grafismo*, contudo, só será introduzida no domínio das relações visuais, se não forem seguidos os três princípios citados [(BERTIN, 1967) e (MARTINELLI, 1991)].

Todo esse processo de comunicação cartográfica, que tem no mapa[41] o seu suporte de informação, oscila entre a capacidade mental humana de interpretar a miscelânea de dados e informações de campo, com limitações de ordem quantitativa e a eficiente capacidade de processamento desses dados e informações por um sistema computacional, com limitações de ordem qualitativa (este trabalho tem como uma de suas metas abrandar esta limitação).

1.5.2. O paradigma dos quatro universos: traduzindo a informação geográfica para o computador

Para tratar do problema fundamental da CIGeo - o entendimento das representações computacionais do espaço -, é necessário um arcabouço conceitual para assimilar o processo de tradução do MR para o ambiente computacional, expresso pelo *paradigma dos quatro universos* de J.M. Gomes e L. Velho, citados em CÂMARA (2002), que distingue:

---

[41] Mapa, nesse contexto, tem um sentido que vai muito além de uma folha de papel repleta de símbolos, linhas e pontos.

- O *universo do mundo-real* (MR), que inclui as entidades da realidade a serem modeladas no sistema: tipos de solo, dados geofísicos e topográficos, etc. (V. Figura 1.4).
- O *universo conceitual* (matemático), que inclui uma definição matemática (formal) das entidades a serem representadas; aí distinguem-se as grandes classes de dados geográficos (dados contínuos e objetos individualizáveis) e, também, pode-se especificar essas classes pelos tipos de dados geográficos utilizados: dados temáticos e cadastrais, modelos numéricos do terreno, dados de detecção remota por satélite, etc.

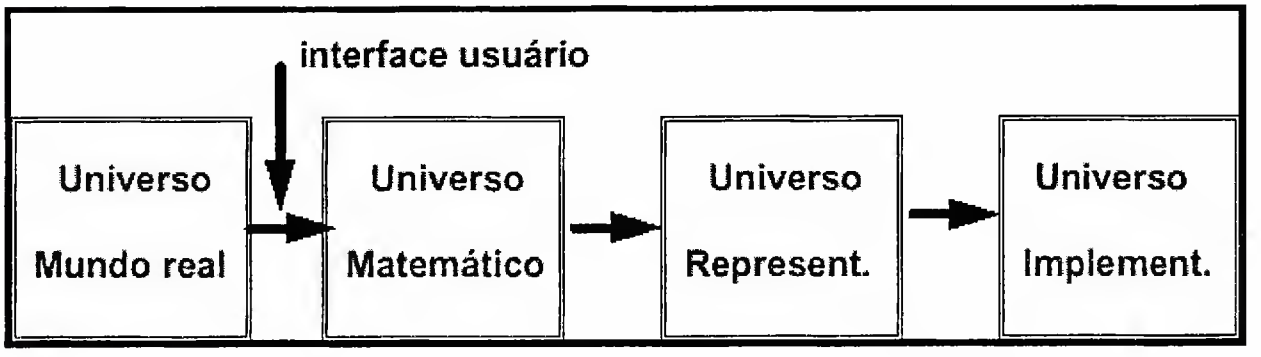

Figura 1.4: Paradigma dos quatro universos[42]

- O *universo de representação*, em que as diversas entidades são traduzidas para representações geométricas e alfanuméricas no computador, que podem variar conforme a escala, a projeção cartográfica escolhida e a época de aquisição do dado; é o domínio das representações matricial e vetorial.
- O *universo de implementação*, em que as estruturas de dados e algoritmos são escolhidos, fundando-se em considerações como desempenho, capacidade do equipamento e volume de dados, para implementar as *geometrias* do universo de representação; é neste nível que acontece a codificação e a realização do modelo de dados por intermédio de linguagens técnicas de programação (LTPs).

O paradigma não se restringe apenas aos sistemas dedicados ao *Geoprocessamento*. Na verdade, o paradigma já vinha sendo aplicado de forma generalizada em disciplinas mais antigas, como a Computação Gráfica e o Processamento Digital de Imagem.

Com pouca variação, pode-se perceber que a horizontalidade das transformações da Figura 1.4 pouco difere em teor da verticalidade das da Figura 2.1, havendo um verdadeiro

---

[42] Extraída de CÂMARA (2002).

consenso nessa visão, também partilhada por BERNHARDSEN (1992), quando diz que: *"Um SIG representa uma visão simplificada do MR"* (V. Figura 1.5).

A TMO acomoda muito bem essa visão de consenso dos autores em tela, ao aproveitar o conhecimento de que os humanos pensam em termos de objetos. A capacidade de abstração permite ao homem ver o todo em vez das partes componentes; é possível pensar numa praia no lugar de grãos de areia e, numa casa, em vez dos tijolos e telhas que a compõem (DEITEL, 2000). É daí que vem toda a base do progresso científico, muito bem incorporada à Ciência da Computação, que é a capacidade que o ser humano tem de abstrair, ou seja, de observar dois ou mais objetos, esquecer as suas possíveis diferenças aparentes, concentrar-se em algumas de suas semelhanças e vice-versa (GUIMARÃES, 1985).

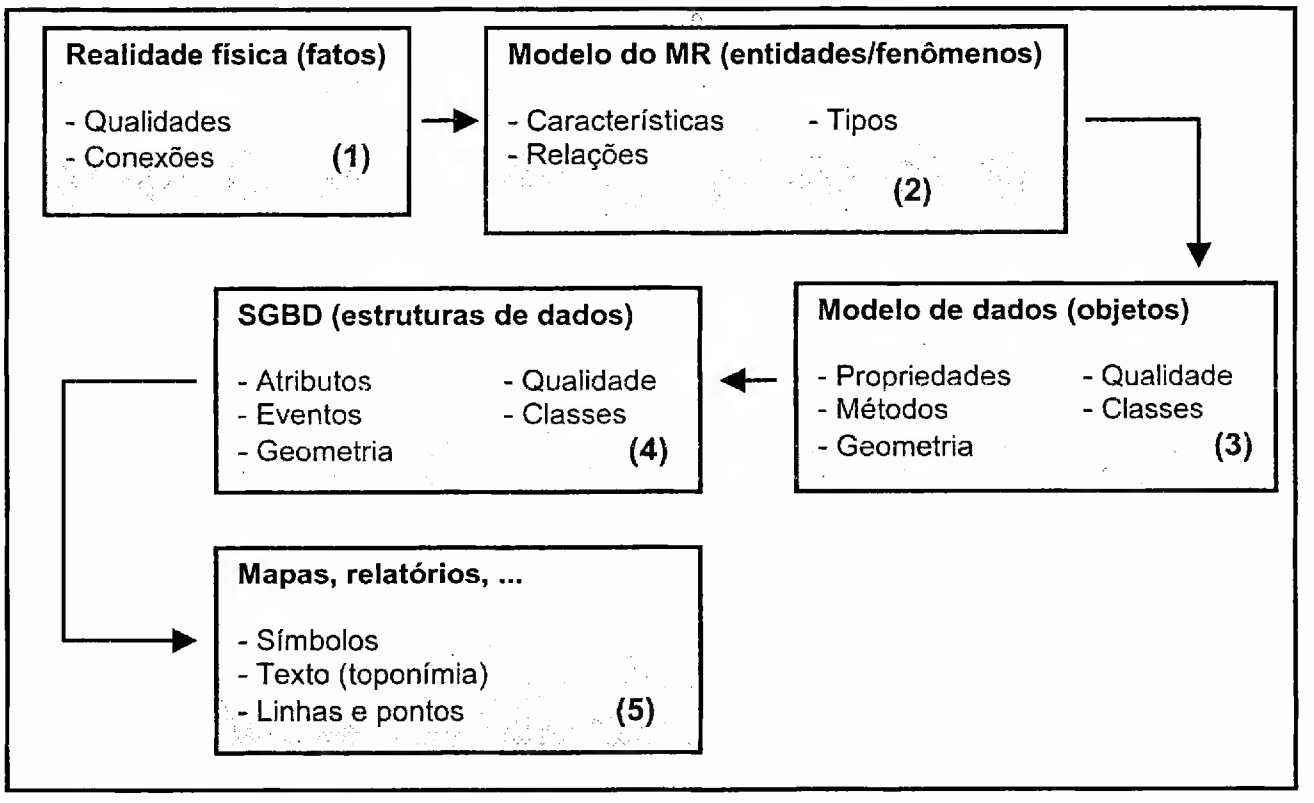

Figura 1.5: Como se insere o MR num SIG por meio de modelos (simplificações)[43].

Após o processo de *abstração*, vem a *representação* dessa abstração (GUIMARÃES, 1985). No caso desta tese, dois dos instrumentos utilizados para avaliar a similaridade semântica entre classes de objetos espaciais (os protótipos denominados de PROFAX e

[43] Adaptada de BERNHARDSEN (1992).

PRONTO©[44]), também constituem formas de representação da capacidade de julgamento de semelhanças e diferenças entre esses objetos.

Deve-se atentar para que a interface com o usuário de um SIG deve, tanto quanto possível, reflete o universo conceitual e esconda pormenores dos universos de representação e implementação. No nível conceitual, o usuário lida com conceitos mais próximos da realidade e minimiza a complexidade envolvida nos diferentes tipos de representação geométrica (CÂMARA, 2002).

1.5.2.1. O universo do mundo-real (MR)

Um aspecto central do *Geoprocessamento* é percebido na característica lógico-matemática dos sistemas de informação. Para que a IG possa ser representada em ambiente computacional, é preciso associá-la a uma escala de medida e de referência, que será utilizada pelo SIG para caracterizar o tipo de IG em uso.

A escala de mensuração utilizada permite associar grandezas numéricas a cada objeto a ser representado no computador. Esse enfoque deriva do conceito de *representatividade*, proposto pelo filósofo Bertrand Russell: *"As propriedades não são intrínsecas aos objetos, mas são obtidas com base em medidas"*. Assim, a representação de um objeto geográfico num SIG dependerá da escala de uso.

As regras de mensuração do fenômeno geográfico determinam o seu nível e cada nível de medida descreve a entidade de estudo com um determinado grau de especificação, que varia de informações quantitativas até informações qualitativas. Como a forma de se medir as variáveis do mundo-real afeta seus modos de manipulação, é essencial que o nível de medida utilizado seja incorporado a um conjunto de observações.

À semelhança da Probabilidade e Estatística, foram propostas quatro escalas de mensuração para o fenômeno geográfico: *nominal, ordinal, intervalar* e *de razão*.

Os dois níveis *nominal* e *ordinal* são temáticos, porque a cada medida é atribuído um número ou nome que associa a observação a um tema ou classe.

O nível de medida *ordinal* caracteriza-se por atribui valores ou nomes para as amostras e gera um *conjunto ordenado de classes*, com base em critérios específicos.

Uma característica importante dos níveis de *medidas temáticas* é que elas *não determinam magnitude*.

---

[44] PROtótipo de avaliação de similaridade semântica entre classes de entidades espaciais, representadas numa ONTOlogia *ad-hoc*.

Quando o estudo necessita de uma *descrição mais específica*, que permita comparar intervalo e ordem de grandeza entre eventos, recorre-se aos níveis de medidas denominados de *numéricos* (por *intervalo* e por *razão*), em que as regras de atribuição de valores baseiam-se numa escala de números reais.

As medidas temáticas e as numéricas de intervalo não devem ser usadas diretamente em expressões matemáticas. Entretanto, na prática, os modelos ambientais combinam valores de razão com valores intervalares. Nesses casos, parâmetros devem ser incluídos para permitir a conversão de valores medidos no nível intervalar para o nível de razão, em unidades apropriadas (CÂMARA, 2002).

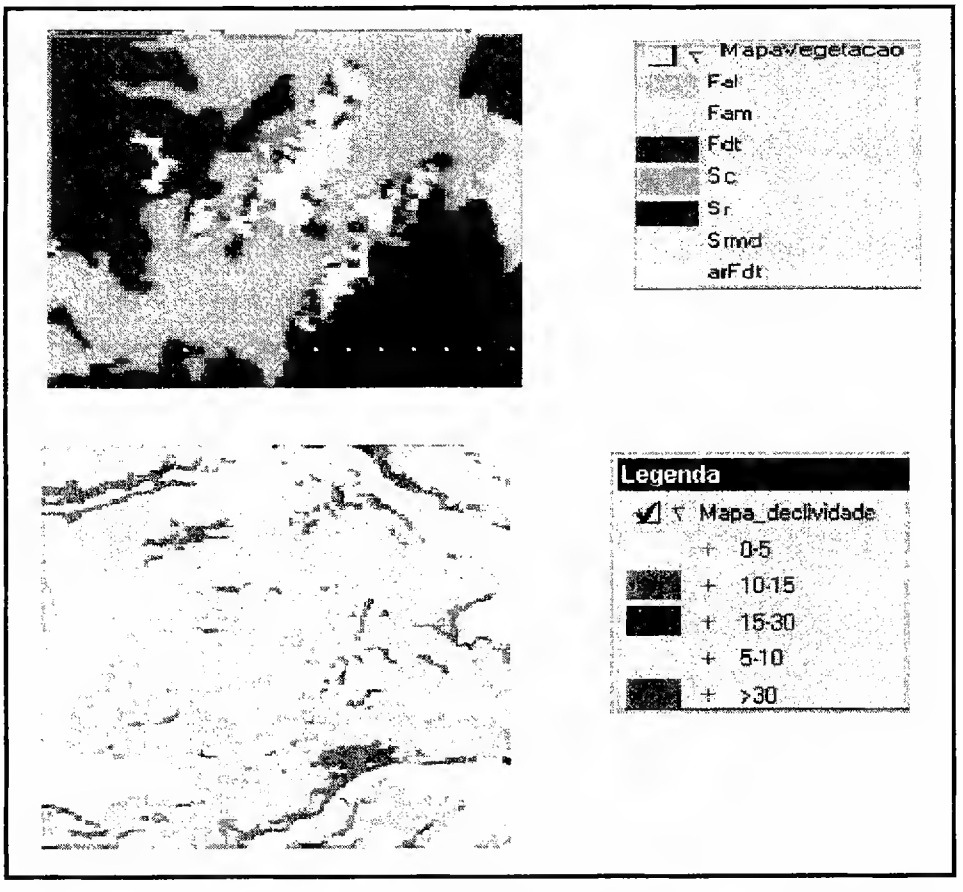

Figura 1.6: Mapa de vegetação (de cima) e mapa de declividade (inferior)[45].

1.5.2.2. O universo conceitual

Antes de entrar na apreciação deste universo, é necessário classificar os *tipos de dados* utilizados no *Geoprocessamento*, quais sejam: os *temáticos*, os *cadastrais*, os de *rede*, os de modelo digital do terreno (*MDT*) e os de *imagem* (CÂMARA, 2002).

---

[45] Extraída de CÂMARA (2002).

Dados temáticos: descrevem a distribuição espacial de uma grandeza geográfica, expressa de forma qualitativa, como os mapas de pedologia e de aptidão agrícola de uma região. Esses dados, obtidos com base no levantamento de campo, são inseridos no sistema por digitalização ou por classificação de imagens. Os mapas da Figura 1.6 (um mapa de vegetação e um de declividade do terreno) foram construídos com dados temáticos.

Dados cadastrais: distinguem-se dos temáticos porque são instanciáveis, i.e., individualizáveis em objetos geográficos, que possuem atributos e podem estar associados a várias representações gráficas; p.ex: os lotes de uma cidade são elementos do espaço geográfico que possuem atributos como proprietário, localização, valor venal, IPTU devido, etc., os quais podem armazenados num SGBD.

Figura 1.7: MDT de uma região montanhosa da Nova Guiné.

Em *Geoprocessamento*, o conceito de **rede** denota as informações associadas a serviços de utilidade pública, como água, luz e telefone, redes de drenagem e rodovias, p.ex. No caso do dado de rede, cada objeto geográfico (cabo telefônico, cano d'água) possui uma localização geográfica exata e está sempre associado a atributos descritivos, presentes no banco de dados.

O termo **MDT** é utilizado para denotar a representação quantitativa de uma grandeza que varia continuamente no espaço. Um MDT, em geral, vem sempre associado ao relevo (V. Figura 1.7), mas podem ser utilizados para descrever outros fenômenos, como a distribu-

ição de calor, campos de força, etc. Entre as aplicações de MDT enumeram-se as seguintes, citando BURROUGH (1987):

- Armazenamento de dados de altimetria para gerar mapas topográficos;
- Análises de corte e aterro para projeto de estradas e barragens;
- Avaliação de mapas de declividade e de vistas perspectivas, para apoio a análises de geomorfologia e erosão;
- Análise de variáveis geofísicas e geoquímicas;
- Apresentação tridimensional do fenômeno (relevo, superfícies diversas, etc.).

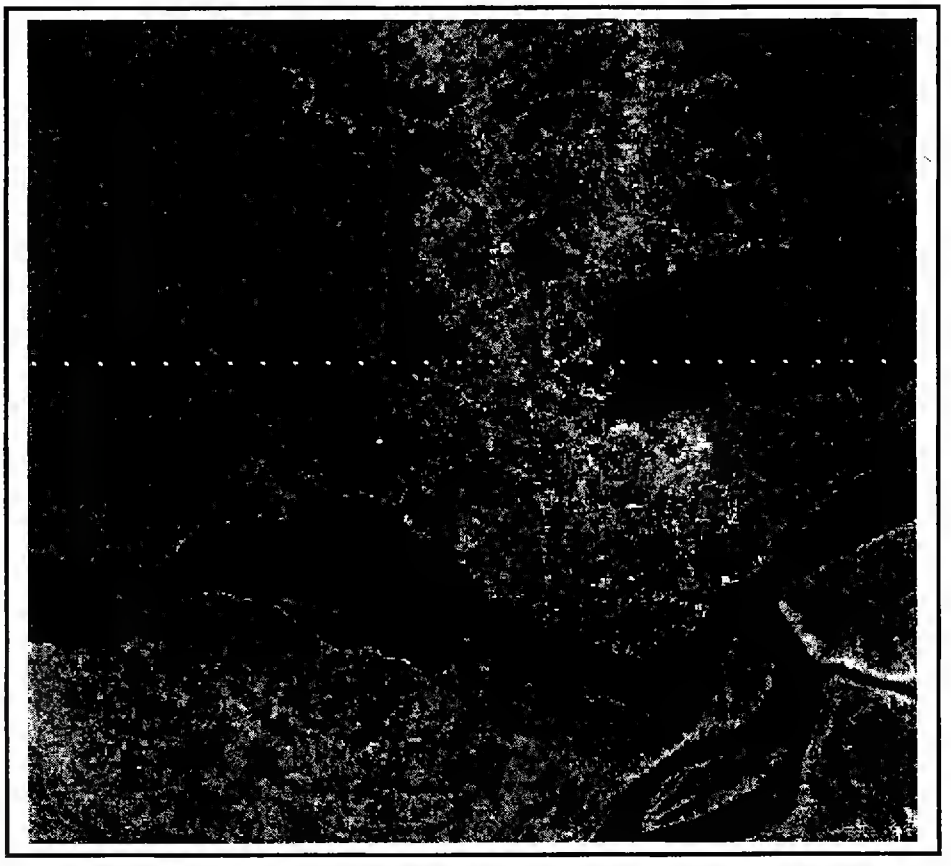

Figura 1.8: Imagem colorida da região de Manaus (AM), tomada pelo satélite TM-Landsat.

Mais especificamente, um **MDT** pode ser definido como um modelo matemático que reproduz uma superfície real, valendo-se de algoritmos e de um conjunto de pontos, cada um determinado por duas coordenadas referenciadas a algum sistema de posição. A cada par ordenado que representa um ponto no plano é associada uma terceira coordenada, que

representa altura ou profundidade, descrevendo a variação contínua da superfície no espaço tridimensional[46].

Obtidas por satélites ou de sensores aerotransportados, as **imagens**[47] (V. Figura 1.8) representam formas de captura indireta de informação espacial. Armazenadas como matrizes, cada elemento de imagem (*pixel* – V. glossário) tem um valor proporcional à energia eletromagnética refletida ou emitida pela área da superfície terrestre correspondente. Os objetos geográficos contidos numa imagem são obtidos por técnicas de fotointerpretação e de classificação, para que sejam individualizados.

É com base nesses conceitos de dados geográficos que um usuário das tecnologias de *Geoprocessamento* (SIG, p.ex.) é capaz de representar o espaço geográfico do seu interesse e sobre ele realizar cálculos e análises qualitativas.

CÂMARA (2002) distingue duas visões complementares de modelagem para o espaço geográfico: o *modelo de campos* e o *modelo de objetos*.

Pelo modelo de *campos*, "enxerga-se" o espaço geográfico como uma *superfície contínua*, sobre a qual variam os fenômenos a serem observados, segundo diferentes distribuições (V. Figura 1.6).

O modelo de *objetos* representa o espaço geográfico como uma coleção de objetos distintos e identificáveis; p.ex: um cadastro espacial dos lotes de um município, os rios de uma bacia hidrográfica ou os aeroportos de uma região, todos com seus atributos identificadores.

Seguem-se os passos para se definir um *modelo conceitual completo* do espaço geográfico:

1º) *Definir as classes* básicas do modelo e estabelecer as suas relações, dentro dos princípios de especialização, generalização e agregação;

2º) Com base no modelo, *definir um esquema* conceitual para um banco de dados geográfico, por especialização das classes básicas.

Nesse tópico, cabe assinalar algumas noções que tomam parte do conceito de *modelo de espaço geográfico*; são elas: *região geográfica*, *campo geográfico* (*geocampo*) e *objeto geográfico* (*geo-objeto*).

Define-se uma *região geográfica R* como uma superfície qualquer pertencente ao espaço geográfico, que pode ser representada num plano ou reticulado e que depende de uma projeção cartográfica. A região geográfica serve de suporte geométrico para localização de

[46] A dissertação de BORGES (1993) entra em pormenores sobre esse tipo de dado.
[47] Imagens: são muito comuns as cenas de satélites e as fotografias de aeronaves (fotos aéreas).

entidades geográficas, porque toda entidade geográfica será representada por um ponto ou um conjunto de pontos em *R*.

Um *campo geográfico* representa a distribuição espacial de uma variável que possui valores em todos os pontos pertencentes a uma região geográfica, num dado tempo *t*. Dessa maneira, torna-se possível representar a variação cronológica de alguns temas (uso do solo, variação climática, etc.).

Um *objeto geográfico* é um elemento único que possui atributos não-espaciais e está associado a múltiplas localizações geográficas. A localização deve ser exata e o objeto distinguível de seu entorno. Três são os desdobramentos dessa definição:

- As *projeções cartográficas:* a esfericidade da Terra é transformada numa projeção plana, subdividida ou recortada arbitrariamente em quadriláteros (quadrículas), que podem dividir a localização de um objeto contínuo, como um rio, por exemplo.
- *Representações geométricas em diferentes escalas:* na prática, num mesmo banco de dados geográfico, podem coexistir representações da mesma realidade geográfica, em diferentes escalas.
- *Múltiplas representações temporais:* as diferentes representações de um mesmo objeto podem corresponder às suas variações temporais, como no caso de um lago que teve suas bordas alteradas (Mar de Aral, entre o Cazaquistão e o Uzbequistão).

Em muitas situações é conveniente permitir a associação de informações não-espaciais a um banco de dados *georreferenciados*. Assim, define-se um *objeto não-espacial* (toponímia, p.ex.) como um objeto que não possui feições, traços ou características espaciais associadas, o que acaba por englobar qualquer tipo de informação que não seja *georreferenciada* e que se queira carregar num SIG.

1.5.2.3. O universo de representação

Neste universo, segundo CÂMARA (2002), definem-se as possíveis representações geométricas que podem estar associadas às classes do universo conceitual. Inicialmente, deve-se considerar as duas grandes classes de representações geométricas: *representação vetorial* e *representação matricial*.

Na *vetorial*, a representação de um objeto é uma tentativa de reproduzi-lo o mais exatamente possível. Qualquer entidade ou elemento gráfico de um mapa é reduzido a três formas básicas: pontos, linhas e polígonos.

A *matricial* consiste no uso de uma malha quadriculada regular[48], sobre a qual se constrói, célula a célula, o elemento que está sendo representado. A cada célula, atribui-se um código referente ao atributo estudado, de tal forma que o programa de computador identifique a que objeto pertence determinada célula.

Vale ressaltar que as representações estão associadas aos tipos de dados anteriormente apresentados (temáticos, cadastrais, de rede, de MDT e de imagem).

1.5.2.4. O universo de implementação

Neste universo, de acordo com CÂMARA (2002), vêm à tona preocupações com as estruturas de dados que deverão ser utilizadas para construir um sistema de *Geoprocessamento*.

É também nesta etapa que são tomadas as decisões concretas de programação, levando-se em consideração os aspectos de *hardware* e de *software* que interferirão no desempenho do sistema.

Um ponto de expressiva relevância a ser levado em conta no universo de implementação é o uso de estruturas de indexação espacial. Os métodos de acesso a dados espaciais compõem-se de estruturas de dados e algoritmos de pesquisa e recuperação e representam um componente determinante no desempenho total do sistema. Em geral, estes métodos são baseados em árvores de busca.

1.5.3. A Inteligência Artificial

Como o espaço-problema (escopo) da IA é um manancial de aplicações de disciplinas de outros campos científicos, por conseguinte, faz-se necessário tentar extrair uma terminologia comum sobre alguns conceitos que serão utilizados nesta pesquisa, a saber: *modelo*, *abstração*, *agente inteligente* (*AI*), *representação do conhecimento* e *ontologia*.

1.5.3.1. Terminologia básica

Os conceitos enumerados acima são essenciais ao estudo do fenômeno da *similaridade semântica* (SS).

A começar por ***modelo***, antes de expor a acepção que lhe é atribuída nessa disciplina, vale a pena examinar as origens e contribuições das quais os pesquisadores da IA se valeram para definir o conceito desse termo.

---

[48] Quadrículas (quadrados) de mesmas dimensões.

CÂMARA (2002), BERNHARDSEN (1992) e BÄHR (1996), nas Figuras 1.4, 1.5 e 2.1, já delinearam definições sobre o termo *modelo*, ficando a idéia de *simplificação* associada ao termo, numa situação de transformação de fenômenos espaciais para um ambiente computacional.

A definição de *modelo* de SETZER (1989), RUMBAUGH (1994), FURLAN (1998) e DEITEL (2000) está vinculada ao desenvolvimento de SGBDs e aplicações *ad-hoc*. Para os quatro autores é unânime a idéia de que um modelo é uma abstração com a finalidade precípua de simplificar uma certa complexidade que tenha sido identificada.

E o que é uma *abstração*?

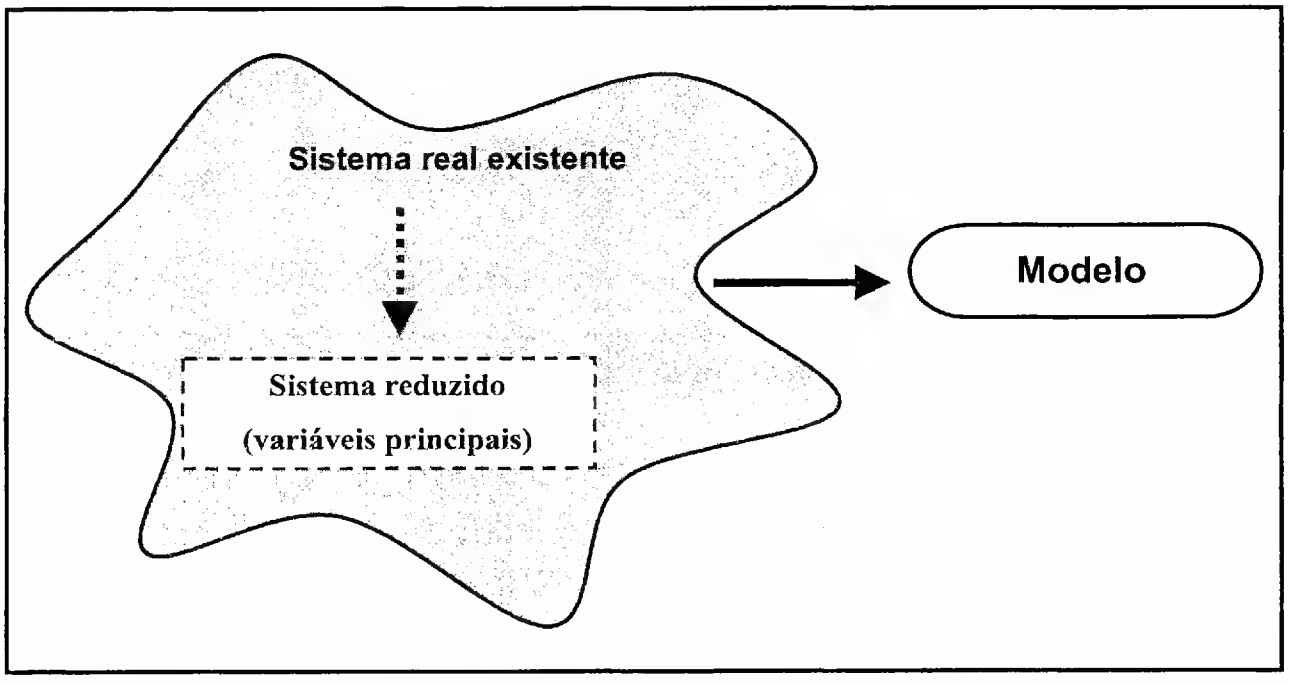

Figura 1.9: Modelagem de um sistema real.

Essa pergunta é central para a Engenharia do Conhecimento[49], que considera três etapas seqüenciais na modelagem da informação: 1ª) *Percepção* do MR, ligada à atenção, sendo de natureza neurológica, interna; 2ª) *Indução* de propriedades, ligada aos mecanismos cerebrais da memória; 3ª) *Abstração* na solução, ligada a processos de raciocínio (de um ser orgânico inteligente ou de uma máquina dotada de IA).

[49] Nas engenharias, esta é a ordem, da mais abrangente para a mais restrita: Engª do Conhecimento, da Informação, de Sistemas, de Requisitos e a de Computação (*Hardware* e *Software*).

O ápice evolutivo das três etapas anteriores, a ***abstração***, é o exame seletivo de determinados aspectos de um problema. O objetivo da abstração é isolar os aspectos que sejam importantes para algum propósito de natureza intelectual e suprimir os que não sejam.

Na construção de modelos, portanto, não se deve procurar a verdade absoluta e sim a adequação a algum propósito. Não há um único modelo correto para uma situação, apenas modelos adequados e inadequados. A construção de modelos nas engenharias e na IA é a aplicação da objetividade mais extrema do princípio da *Navalha de Ockham* (V. glossário).

A Figura 1.9 representa a simplificação que um modelo realiza. O sistema real é um conjunto complexo de variáveis, de forma não muito definida, que necessita ser reduzido a um conjunto sobre o qual possam ser aplicadas as regras formais da Matemática ou as fórmulas empíricas da Engenharia. O sistema reduzido é o núcleo do sistema real existente, que basicamente "dita" o comportamento deste e que pode ser modelado para ser submetido à análise, dentro de uma estrutura conhecida. A Pesquisa Operacional é uma disciplina recentemente surgida e que explora muito esse conceito de modelo.

SETZER (1989) teoriza a *abstração* de informações e dados em cinco níveis: *1º) Real; 2º) Informações informais; 3º) Informações formais; 4º) Dados; e 5º) Máquina.* Do primeiro para o quinto, diminui a complexidade e aumenta o formalismo.

O *primeiro nível* (mundo-real - MR) é nebuloso. Foco de estudos de correntes filosóficas dos materialistas e dos idealistas, é o mundo dos seres, dos fatos, das coisas, dos organismos sociais. Até bem recentemente, antes da ascensão das visões ontológicas, técnicas muito restritivas de análise e projeto é que determinavam o que era de interesse do MR para ser formalizado nos níveis inferiores.

O *segundo nível* é caracterizado pela linguagem natural, carregada de muita ambigüidade. Muito se discute sobre a capacidades e limitações dessa forma de comunicação nas ciências cognitivas, mas o que importa na visão sistêmica é que esse já é um nível de abstração e de muita generalidade. O espectro lingüístico nesse nível vai desde uma asserção logicamente perfeita até frases ambíguas (poesia), contendo um simbolismo que transcende a experiência sensorial direta (metáforas). É também chamado de nível descritivo. O *modelo descritivo* é a simplificação abstraída da realidade que se faz nesse nível.

O *terceiro nível* é caracterizado pela construção de um modelo formal, porque servirá de base para implementar programas no nível mais baixo - o de máquina ou computacional. Qualquer erro aí cometido é potencializado para as camadas inferiores de abstração. É também chamado de nível de *modelo conceitual*, uma vez que o sistema de conceitos é rigoroso, embora ainda transcenda o formalismo matemático.

O *modelo conceitual* possui duas estruturas intrínsecas que o caracterizam: estruturas de *informações* e estruturas de *manipulação* das informações. A primeira trata das entidades e dos relacionamentos entre as entidades. Na verdade, essas estruturas são *metainformações* (informações que descrevem as informações). A segunda é mais funcional e trata de operações de manutenção e controle das entidades do modelo: definir, atualizar, eliminar e ler as *metainformações*. Nesse nível de modelo conceitual, já existem linguagens para especificar as *estruturas* das informações, denominadas de LDEI (linguagens de definição das estruturas de informação) e para especificar as *manipulações* de informações (LMI). Esta tese explora muito os *conceitos* desse nível, no que tange à informação geográfica.

O *quarto nível* é o nível dos símbolos que serão introduzidos no computador: os *dados*. É também chamado de *nível operacional*. Há uma distinção fundamental entre informações formais e dados: as primeiras podem seguir qualquer formalismo matemático, existindo tanto no papel como na mente humana. Os dados, não. Estes devem ser representados de tal forma que um computador possa recebê-los e tratá-los.

Segundo SETZER (1989), já é história a *amigabilidade* de sistemas operacionais[50] dos computadores, responsável por fazer com que os dados cheguem cada vez mais próximos das informações formais. O usuário não precisa mais aprender "programês" para usar o computador. O que era sonho já está se tornando realidade, porque, numa escala gradual, o usuário que sabia "programês", já livre desse fardo, passou para a era do "analistês", sabendo usar alguns recursos de linguagens gerais de especificação de sistemas, podendo criar sistemas especialistas. Analogamente ao nível conceitual, há no nível de dados linguagens de *descrição* de dados (LDD) e as de *manipulação* (tratamento) de dados (LMD, p.ex: as *query languages* – SQL).

O *quinto nível* é o de *máquina*, sobre o qual não vale a pena estender uma descrição pormenorizada nessa altura da pesquisa.

Pode-se perceber que o enfoque de modelo de SETZER (1989), muito bem adaptado à modelagem de bancos de dados, é bem similar aos vistos nas Figuras 1.4 e 1.5.

FURLAN (1998), mais detido na OO, não fogiu à regra de criar e utilizar modelos simples e genéricos para resolver problemas que surgem da observação de fenômenos do MR. Como boa prática de abstração, o autor ressaltou que esses modelos devem fundamentar-se na essência e não nos efeitos aparentes dos fenômenos, o que lhes renderá mais resistência ao teste do tempo, fazendo com que a implementação e a manutenção dos sistemas que deles se utilizem saiam menos custosas.

Neste ponto, um parênteses. Nada como citar as críticas epistemológicas de SÖRGEL (1999) sobre o desenvolvimento da Ciência da Computação (OO) e da Ciência da Informação, no que tange às teorias e técnicas de classificação e recuperação da informação.

Dagobert Sörgel colocou a ontologia num quadro conceitual que o autoriza a assemelhá-la a uma classificação, embora com mais condições de responder a funções críticas de SRIs e de sistemas de aprendizagem de máquina (domínios da IA). Essas condições dotam essa estrutura do conhecimento de características que ainda não foram totalmente exploradas pela engenharia de sistemas e pelos cientistas da computação, mas que vêm há séculos sendo discutidas por filósofos (Aristóteles, séc. IV A.C) e sistematizadas desde o séc. XVII[51] até os dias de hoje por bibliotecários e cientistas da informação.

Outro ponto que precisa de esclarecimento prévio, porque pode tornar-se motivo de polêmica quando vier à tona em diversos trechos do trabalho, é o uso do termo "ciências da classificação", utilizado por VICKERY (1980), por SÖRGEL (1999) e por MODELL (2001) para designar um grupo de ciências (Biblioteconomia, Ciência da Informação e Arquivologia, p.ex.) ou áreas de estudos interdisciplinares que têm nos princípios lógicos de classificação documentária a base de sua metodologia.

D. Sörgel não desejou reclamar direitos de autoria para as ciências da classificação sobre o "descobrimento da roda". Aplausos para os cientistas da computação que engendraram a OO, talvez com pouquíssima ou nenhuma contribuição da Terminologia ou da Biblioteconomia. Mas será que se houvesse uma parceria maior entre esses campos, a AOO (análise orientada a objeto) não teria deslanchado mais cedo? Não teria assumido uma estrutura mais robusta em termos conceituais, poupando precioso tempo na orladura da atualmente tão aclamada UML™?

É este o fulcro da questão para D. Sörgel, que considera uma repetição de erros, na área da IA, o desenvolvimento de sistemas como os projetos CYC©[52] (construção de extensas bases de conhecimentos, orientadas por ontologia) e *Wordnet*™ (taxinomia *on-line*), ambos sendo executados sem uma suficiente participação de especialistas em construção de *tesauros* e classificações. Não se trata de reconhecer quem inventou a roda, mas por que motivo replicar esforços em reinventá-la?

Vale a pena, também, averiguar as fontes de conhecimento sobre o que seja *modelo* na *Computação Gráfica*. Daí, contribuições de igual valor dos trabalhos clássicos de NEW-

---

[50] *User-friendly operational systems.*
[51] O clássico *Advis pour dresser une bibliothèque*, de Gabriel Naudé, 1644 (JANNUZZI, 2002, p.3).
[52] RODRÍGUEZ (2000, p.25) descreve este projeto.

MAN (1979), HARRINGTON (1983), FOLEY (1984) e ROGERS (1985). Esta síntese de definições de modelo vai solidificar a definição proposta na área de IA (a seguir), muito importante na elaboração de um agente inteligente (AI).

Para os autores da Computação Gráfica, um modelo só tem sentido se houver uma aplicação (problema a resolver). Destarte, um *modelo de aplicação* é um componente conceitual de aplicações gráficas interativas, que incorpora a descrição de objetos e de estruturas de dados para representar entidades (físicas ou não) e fenômenos, não apenas com propósitos gráficos (cenas, imagens e diagramas), mas, em geral, para representar a essência e o comportamento dessas entidades e fenômenos.

Percebe-se da definição anterior que os modelos não são exclusividade das ciências exatas. Pesquisadores das ciências sociais também têm neles um manancial de ferramentas para as simulações, testes de hipóteses, predição do comportamento de entidades e fenômenos como a compreensão, visualização e aprendizagem.

Os modelos também ajudam no entendimento de sistemas complexos, compostos de vários subsistemas que interagem entre si, bastando alterar o conjunto de parâmetros de entrada em cada subsistema e verificar os efeitos de saída no sistema como um todo.

Enfim, sintetizando o enfoque da Computação Gráfica, um modelo é uma rica descrição dos componentes e processos que especificam tanto a *estrutura* como o *comportamento* de uma entidade ou fenômeno do MR e que estão sendo representados (modelados).

Entrando na IA, o conceito de modelo, além de se servir do repertório das definições anteriores, incorpora reflexões da Filosofia, especialmente aquelas ligadas à corrente do *utilitarismo*, adaptado da escola *epicurista*[53] pelos filósofos iluministas do século XVIII. Para os epicuristas, prováveis fundadores da Ética, a máxima era: "Prazer pela prática da virtude e do bem". Já a adaptação dos iluministas para o *utilitarismo* resumia-se ao seguinte tema nuclear: "Saber viver bem; buscar o saber que permita acertar" (BESSE, 1998).

A definição de modelo na IA, segundo RUSSELL (1995), introduz de forma cristalina o princípio utilitarista na construção de um sistema que age racionalmente (um agente inteligente ou AI).

Na IA, o modelo pertence ao mundo da Lógica, preferencialmente da Lógica de Primeira Ordem (LPO)[54], mais expressiva do que a Lógica *Proposicional* (LP). Esta última admite o mundo composto apenas de fatos e não possui uma notação gráfica tão rica quanto à da LPO. Esta reforça a representatividade do mundo, porque o admite não só composto de fa-

[53] De Epicuro, filósofo pós-socrático do século IV A.C.
[54] Criada por Gottlob Frege (1879) e melhorada em termos de notação por Giuseppe Peano (1889).

tos, mas também de objetos e de relações entre esses objetos. Algumas dessas relações são funções quantificáveis, que produzem um único valor como resultado de saída, tendo sido dada uma entrada.

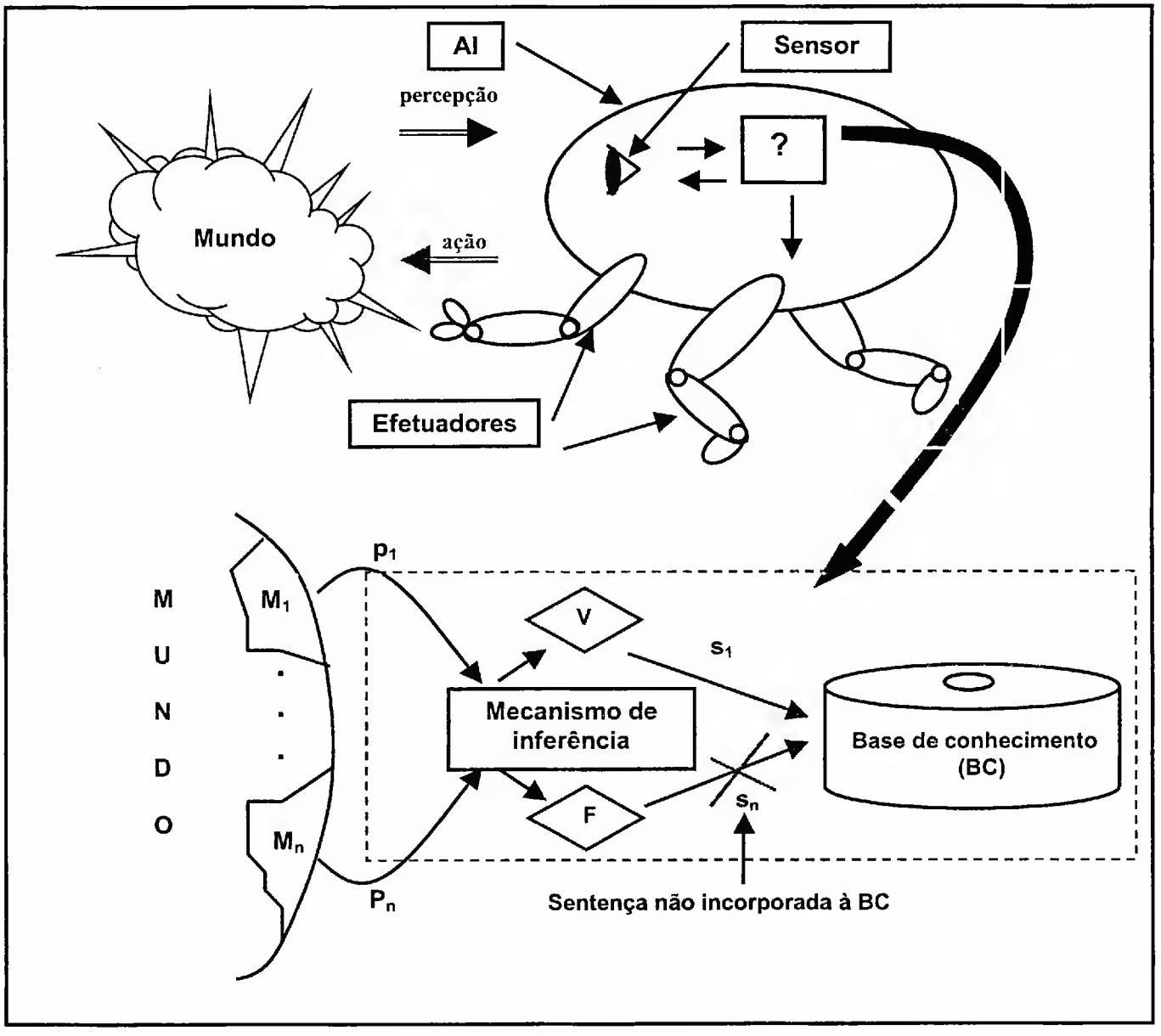

Figura 1.10: Modelo lógico da IA.

Destarte, o *modelo lógico* é um mundo para o qual um argumento é válido segundo uma determinada interpretação. Essa noção de modelo é quase que uma generalização das anteriores, uma vez que, aqui, um modelo não é apenas um conjunto de objetos e estruturas de dados que simplificam a visão do mundo, mas é um mundo (conceitual), criado com base na validade de um argumento que está sendo interpretado (por um homem ou máquina).

Se um argumento contém pouco significado, ou seja, é pouco informativo sobre o mundo, mais modelos serão necessários para suprir a base de conhecimento (BC) e vice-versa.

A Figura 1.10 ilustra a noção de modelo segundo a IA. Na parte superior da figura, verifica-se um AI recebendo percepções (p_1 ... p_n) do mundo (por meio do sensor) e produzindo ações sobre o mundo (por meio dos *efetuadores*), após a avaliação das sentenças lógicas decorrentes das percepções.

Na parte inferior da figura, verifica-se a expansão do quadrado do AI, com uma interrogação que simboliza o mecanismo de inferência e a BC (módulo de validação das partições do mundo). Se uma sentença advinda de uma percepção (p_1) for verdadeira[55] (s_1), ela é validada na BC e desse encadeamento poderá ocorrer uma ação. Se uma sentença advinda de uma percepção (p_n) for falsa (s_n), não haverá ação decorrente desse encadeamento e tal sentença não será incorporada à BC para formar um argumento válido.

É dessa forma que as parcelas do mundo *factual* (M_1 ... M_n) integram a BC do AI por simbolismo lógico, de tal sorte que se essa base contiver todos os argumentos válidos que formalizam o mundo, não será mais necessário haver partições do mundo sobre as quais o AI deva inferir (tirar conclusões, raciocinar).

Dizendo de outra forma, na parte inferior da figura, verifica-se o mecanismo de validação das partições do mundo, em que uma delas (M_1 ... M_n) só pode ser incorporada à BC do AI, caso o mecanismo de inferência deste AI relacione a partição à cláusula lógica de validade (V). Assim relacionada (por meio de cálculo formal), a partição transforma-se num objeto lógico-matemático denominado de *modelo lógico*.

As fontes de reflexão que influíram nas atuais concepções sobre AIs vêm dos gregos antigos. Da época da Filosofia Clássica, assinala-se o *Diálogo de Êutifron*, relatado por Platão. No diálogo, Sócrates perguntou a Êutifron o que *caracteriza* a lealdade, de tal sorte que se pudesse *classificar* atos leais, para se obter um *padrão* de julgamento para as ações humanas. O diálogo pode ser considerado como a descrição de um algoritmo ou um *proto-agente inteligente*, capaz de distinguir um ato leal de um desleal.

Dos clássicos, vieram os pós-socráticos (séc. IV A.C.) e mais contribuições do utilitarismo do séc. XIX, como já foi mencionado. Não se deve perder de vista o princípio básico que deve presidir o comportamento de um agente racional: "Fazer a coisa certa".

Agir (comportar-se) ou *pensar* racionalmente? Esses verbos resumem os problemas básicos da IA, dos quais decorre a própria definição de AI.

---

[55] V. losango de decisão, na figura: V = verdadeiro; F = falso.

Segundo RUSSELL (1995): "A IA é um campo de estudos da Ciência da Computação, preocupado com a automação do comportamento inteligente (AI); por extensão, a IA está preocupada com o projeto de um AI ou de um agente baseado em conhecimento".

Um AI é algo (coisa) que vive num mundo dinâmico, percebendo fatos e agindo (reagindo) nesse ambiente. As percepções são captadas pelos sensores e as ações são realizadas pelos *efetuadores*. Essencialmente, um AI é composto de um *módulo estrutural* e de um *módulo de programa*.

O módulo estrutural subdivide-se em partes destinadas à percepção (sensores), à ação (*efetuadores*) e ao armazenamento de regras (base de conhecimento – BC). A BC é a representação de um conjunto de fatos sobre o mundo. A representação é materializada por sentenças em língua natural, que por sua vez podem ser representadas por uma linguagem de representação do conhecimento (LRC), de natureza simbólica. LRC é uma coleção de princípios da Lógica Proposicional (LP) ou da Lógica de Primeira Ordem (LPO), capaz de ser a portadora do raciocínio dedutivo de um AI.

Há LRCs apropriadas para a IA (LISP, PROLOG e outras), cuja característica fundamental é a de já conter, em sua sintática e semântica, estruturas tipificadas pela LP ou pela LPO. Há outras linguagens – as LTPs ou linguagens técnicas de programação -, como o C++™ ou *Java*™, que são capazes de oferecer condições para construir rotinas que traduzam as estruturas lógicas da LP ou da LPO (NORVIG, 2002).

O módulo de programa *mapeia* o par percepção-ação segundo as regras da BC do AI. Esse *mapeamento* é realizado no mecanismo de inferência do agente (V. Figura 1.10).

Noutras palavras, um AI é uma *máquina de percepção e ação codificadas* (metas, crenças), parcial ou totalmente, capaz de *resolver problemas contingenciais* (difíceis, porque mudam constantemente) de um domínio muito restrito da realidade. Não se pode expandir em demasia este domínio, caso contrário deparar-se-á com um problema intratável (problema NP – V. glossário)

E o que seria um AI ideal? Dada uma seqüência de percepções, o AI ideal é aquele que deve agir de maneira a produzir o maior benefício possível (*utilitarismo*) ao seu desempenho, de acordo com as evidências contidas em sua BC, i.e., um AI ideal é a aquele que melhor responde aos estímulos vindos do mundo. E o que significa "melhor responde"? Significa produzir uma ação bem-sucedida. E o que significa uma ação bem-sucedida? como quantificá-la? Esse é o objetivo central da IA: "Propiciar meios para a tomada de decisão de um AI, com base no seu mecanismo de inferência (raciocínio), aplicado a uma base de conhecimento.

Percebe-se que representar esse conhecimento faz parte da solução dos problemas em IA. Um AI pode inferir sobre o mundo, se possuir uma BC carregada de argumentos que lhe permitam tirar conclusões válidas e, assim, agir racionalmente.

Há três níveis na construção de um AI: 1°) *Cognitivo*: mais abstrato, ligado à espistemologia (como adquirir o conhecimento e avaliá-lo em termos de sua natureza e da aplicação ou problema); 2°) *Lógico*: ligado à codificação (em forma de sentenças) do conhecimento considerado relevante para o AI; 3°) *De implementação*: ligado à determinação da arquitetura do AI, às injunções de máquina e às estruturas de dados.

A Figura 1.11 mostra os quatro tipos de agentes (sistemas) inteligentes que são do interesse da IA. Três deles (quadrantes vermelhos) ainda estão em bases muito polêmicas, não se podendo criar teorias sobre eles. Já o agente que se *comporta* (age) *racionalmente* (quadrante de fundo azul), possui um quadro teórico que o envolve e até produz alguns resultados práticos.

Formalismo → / Ação ↓	HOMEM	RACIONALIDADE
PROCESSOS MENTAIS	Agentes que pensam como humanos.	Agentes que pensam com racionalidade.
COMPORTAMENTO	Agentes que agem como humanos.	Agentes que agem com racionalidade.

Figura 1.11: Tipos de agentes inteligentes.

O quadro da figura foi montado tendo como características essenciais de ação (linhas do quadro) os *processos mentais* (agentes que pensam) e o *comportamento* (agentes que se comportam). Como características essenciais de formalismo (colunas), o quadro conta com o desempenho dos agentes pelo enfoque do *homem* (como humanos) e pelo enfoque da *racionalidade* (racionalmente), tendo em vista que agir racionalmente é estar de acordo

com o princípio utilitarista de "fazer a coisa certa", ou melhor ainda: "fazer a coisa da melhor forma possível", dependendo das limitações de formalização do mundo com o qual a IA está sempre envolvida.

Há quatro espécies de agentes que se comportam racionalmente, classificados segundo as características do *módulo de programa* do agente:

- Agente-reflexo simples (**ARS**);
- Agente-reflexo com estado interno (**AREI**);
- Agente baseado em meta (**ABM**);
- Agente utilitarista (**AU**).

Do ARS para o AU, aumenta a complexidade do sistema. O **ARS** trabalha mais simplesmente; encontra uma regra que contenha uma condição que satisfaça à situação definida pela percepção e, depois, age de acordo com essa regra.

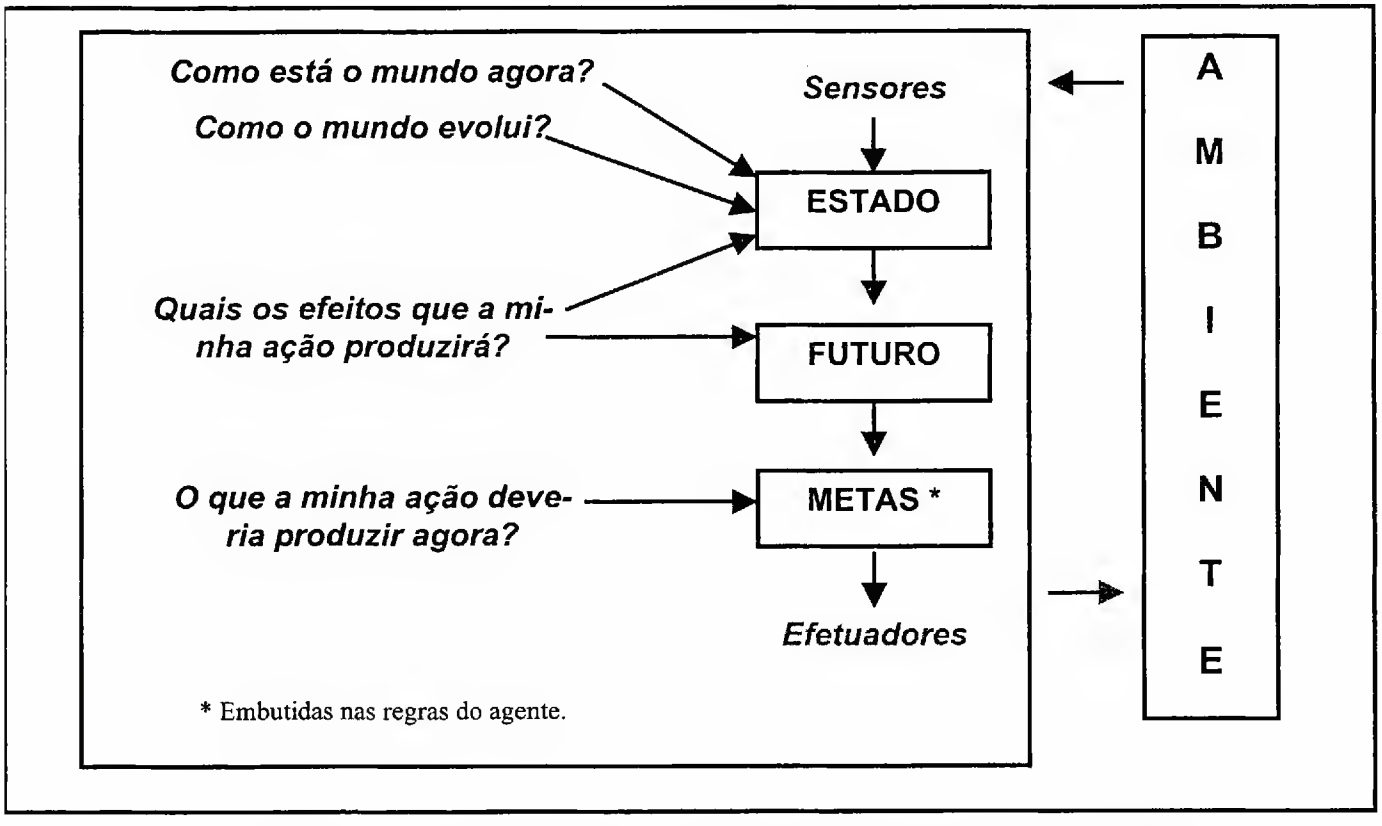

Figura 1.12: Esquema de um agente baseado em metas (ABM)[56].

Neste ponto, cabe uma observação a respeito do enquadramento dos protótipos que serão desenvolvidos neste trabalho. Ambos (PROFAX e PRONTO©) podem ser enquadrados como um agente do porte de um ARS. Como se verá na metodologia, os dois não são

[56] Adaptada de RUSSELL (1995).

estanques e há funções abertas para implementação futura, que poderão proporcionar a evolução do agente para uma das três outras categorias superiores. Se o protótipo mais aprimorado (PRONTO©) receber melhorias, como as propostas para pesquisa futura (V. subitem 7.5.2.6), ele poderá evoluir para um ABM, incorporando mais representatividade ao modelo de avaliação de similaridade semântica que foi estendido neste estudo-de-caso.

O **AREI** é mais aprimorado que o ARS, diferindo deste por determinar a sua ação não apenas de acordo com a regra que contém a situação definida pela percepção imediata sobre o mundo, mas por uma regra mais elaborada, que contém a situação definida pela evolução do mundo com relação a um estado anterior

O **ABM** é um avanço em relação aos dois anteriores. Nesse agente, há um passo adicional de sondagem, que incorpora uma avaliação prospectiva à regra de ação, i.e., sobre os efeitos que esta ação poderia causar no futuro. Em suma, um agente que possui metas e busca por soluções que satisfaçam a tais metas é melhor (mais adequado) do que o agente que apenas reage ao mundo. Em questão de desempenho, os dois agentes anteriores podem superar o ABM, mas este é muito mais flexível que aqueles, o que o torna talhado para aplicações de busca (recuperação) de informação e planejamento.

Por representar algumas das especificações que o PRONTO© deve incorporar para evoluir, um ABM vem ilustrado na Figura 1.12.

Dessa forma, já se poderia introduzir alguns aspectos da metodologia de aprimoramento desse ARS que é o PRONTO©: a ontologia criada para o protótipo pode vir a tornar-se a base de conhecimento do ABM e a linguagem de representação do conhecimento do ABM poderia ter como núcleo de desenvolvimento o código-fonte em *Java*™ que produziu o módulo de edição da ontologia. O mecanismo de inferência do futuro ABM pode evoluir com base no módulo de cálculo de SS, cujas rotinas também foram codificadas em *Java*™. O modelo matemático de RODRÍGUEZ (2000) poderia ser complementado ou alterado segundo outros modelos, como o de SANTOS (2002), GANESAN (2002) ou WONG (2002), mais ligados às tarefas de manutenção e busca de informação (V. subitem 7.5.2.7).

Para encerrar a visualização do futuro do PRONTO© como um ABM, cabe ainda citar uma variação dessa classe de AI: o *agente autônomo.* Segundo RUSSELL (1995), um AI desse tipo não dependeria apenas de uma BC sobre o ambiente para agir e sim de um mecanismo adicional e mais complexo: o da *aprendizagem* pelas observações. Esse mecanismo trabalha com uma função concentrada nos aspectos de indução de experiência, capaz de compensar a ignorância do projetista do AI e da insuficiência de regras de sua BC diante

de um ambiente desconhecido. Tal função pode ser um polinômio, uma rede de crença[57] ou mesmo uma RNA.

O **AU** caracteriza-se por comportamentos racionais mais elaborados que os três anteriores. Um ABM pauta suas ações num domínio binário e determinístico (ou BOM, FELIZ ..., ou RUIM, INFELIZ ...). No caso do AU, surge o conceito de *utilidade: "Função que correlaciona um estado com valores pertencentes ao conjunto dos números reais (ℜ); é um gradiente e não um determinismo binário de satisfação".* O AU é muito apropriado para aplicações de decisão e de jogos, em que o ABM teria dificuldades, p.ex., ao tratar de metas conflitantes como velocidade *vs.* segurança, ou ao se defrontar com um conjunto muito grande de metas, não possuindo uma mensuração de certeza sobre qual das metas seria a mais adequada, numa escala de prioridades para solucionar um problema (RUSSELL, 1995).

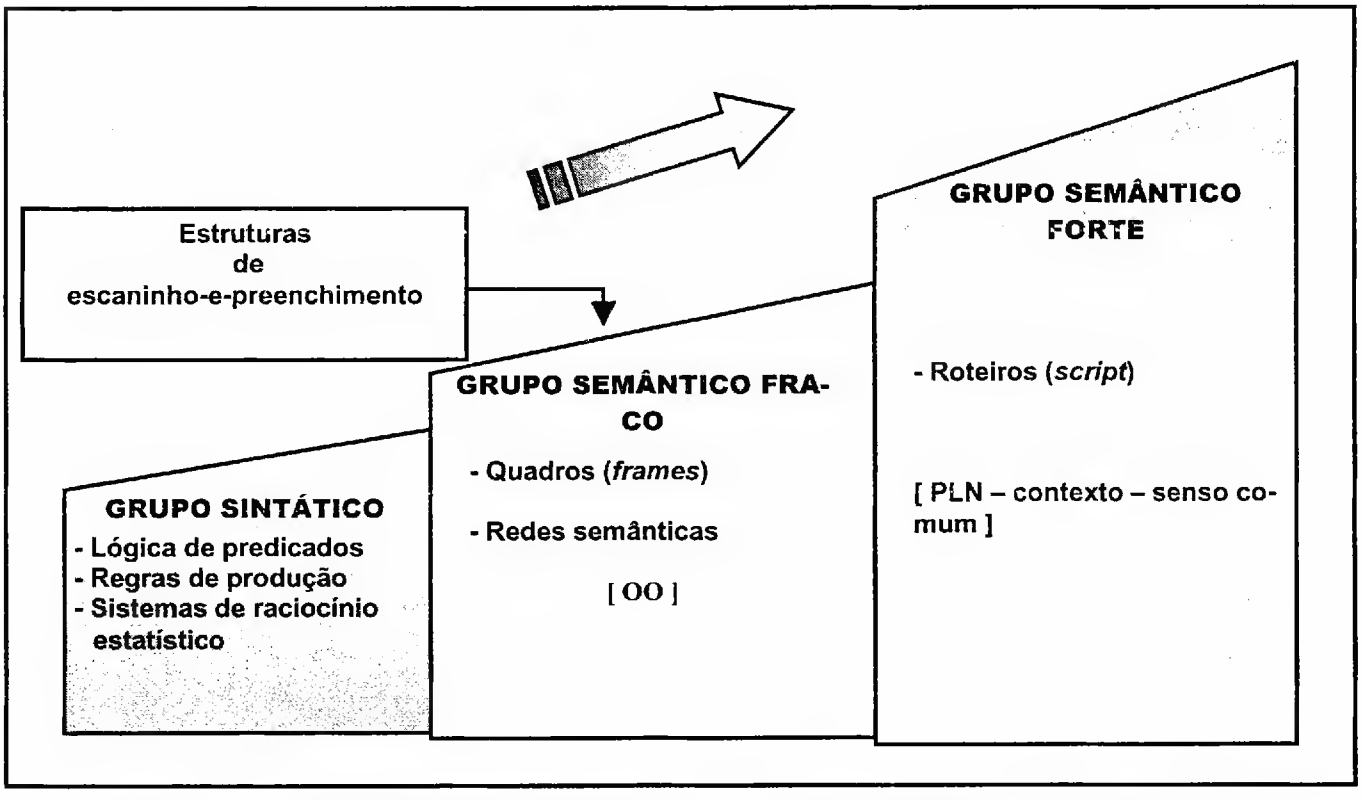

Figura 1.13: Espectro sintático-semântico de representação do conhecimento na IA.

Outro conceito que emana dessa visão geral sobre AIs é o de ***representação do conhecimento*** - expressão do conhecimento numa forma reproduzível em computador, tal que os AIs possam atingir um desempenho razoável se a representação for adequada. O instrumento para essa representação é a LRC (linguagem de representação do conhecimento).

[57] Estrutura de representação da incerteza do conhecimento, baseada numa árvore de distribuição de probabilidades.

RICH (1993) reúne as estruturas de representação do conhecimento em três grandes grupos: o *grupo sintático*, o *grupo semântico fraco* e o *grupo semântico forte*, ilustrados na Figura 1.13. De baixo para cima, na escala, a explicitação (semântica) do conhecimento aumenta.

As *formas* mais *sintáticas* não são adequadas para a representação do conteúdo do conhecimento. Suas regras de inferência são procedimentos que operam sobre fórmulas bem formadas, independentemente do que representam. As *representações lógicas* são típicas desse grande grupo sintático.

As estruturas de *escaninhos-e-preenchimento,* grupo característico das redes semânticas, já estão na área de domínio conceitual do *grande grupo semântico*. As redes semânticas destinam-se a captar os relacionamentos semânticos entre entidades e normalmente são empregadas com um conjunto de regras de inferência que foi especialmente projetado para controlar adequadamente os tipos específicos de arcos presentes na rede (V. Figura 3.9). Os sistemas de quadros (*frames*) são mais estruturados e rigorosos, conceitualmente falando, do que as redes semânticas. As regras de inferência que os quadros comportam são mais especializadas e em número maior que nas redes semânticas.

A *função de herança* é executada de forma muito mais eficiente pelas estruturas de *escaninhos-e-preenchimento* do que pelas estruturas do grupo sintático, porque o conhecimento é representado por entidades, seus atributos e relacionamentos entre tais entidades. Essa vantagem é ampliada pela forte interação com as técnicas de modelagem orientada a objetos (TMO) e pela capacidade que têm de incorporar relações ontológicas de gênero-espécie (*é-um*) e de todo-parte (*é-composto-de*), conciliando aspectos outrora considerados incompatíveis pelas clássicas técnicas de programação estruturada: *modularidade* dos programas *vs.* facilidade de *visualização* das pessoas para o que está por trás do código-fonte.

A Figura 3.9 mostra de maneira inequívoca a verdadeira finalidade das redes semânticas, que é mostrar como o conceito de uma entidade (nó) está relacionado aos de outras entidades. A construção de um sistema de conceitos, segundo FELBER (1984), baseia-se nessa estrutura de forte poder de síntese e de entendimento, muito mais eficiente que a de uma descrição literal das entidades e relacionamentos do sistema.

Encerrando a terminologia básica para conceitos de interesse da IA e que terão aplicação no estudo da SS, seguem-se as mais significativas definições de ***ontologia***, algumas já antevistas em ensaios descritivos e que serão ainda muito reiteradas daqui em diante (subitem 3.2.2.2.3) e mesmo estendidas às raias do experimento, como se examinará no Capítulo 6 deste trabalho.

As seguintes definições são as mais difundidas na literatura, apesar da excessiva generalidade:

"Conjunto de classes amplas de objetos de uma representação." [T. R. Gruber *apud* LEÃO (2003)]

"Avaliação explícita ou representação de uma parte de uma conceitualização." [N. Guarino *apud* RODRÍGUEZ (2000)]

"Entendimento compartilhado em algum domínio de interesse, que pode ser usado como uma infra-estrutura unificada para resolver problemas, permitindo que os resultados de uma pesquisa em um campo possam ser aplicados em outro campo e vice-versa." (USCHOLD, 1996)

LEÃO (2003) consolidou essas e outras definições gerais num entendimento que elucida o papel que as ontologias podem desempenhar no desenvolvimento de SIs. Parafraseando este autor, pode-se definir uma ontologia como um documento (vocabulário, classificação e taxinomia) que explicita os conceitos e relações em algum domínio, bem como o conjunto de axiomas que restringem a sua interpretação. O conteúdo desse domínio do conhecimento é representado por uma especificação menos formal que um modelo conceitual, capaz de propiciar uma compreensão mais fácil da realidade que um determinado grupo social compartilha em termos de experiência. Dessa forma, a ontologia contribui mais efetivamente para eliminar a confusão conceitual e estabelecer uma terminologia comum, sem ambigüidades, para facilitar a comunicação, a representação e o compartilhamento do conhecimento entre os diferentes profissionais de uma organização.

Definições mais específicas e demais considerações sobre ontologias serão examinadas no Capítulo 2 (subitem 2.2.3) e no Capítulo 3 (subitens 3.2.2.2.1 e 3.2.2.2.3). As aplicações dessas definições serão vistas ao longo de todo o Capítulo 6.

1.5.4. As inteligências

Como já discutido no subitem 1.2.2 e anunciado no subitem 1.4, este espaço foi reservado para algumas considerações de ordem epistemológica sobre as ciências que tentam incorporar como seus objetos de estudo um assunto ainda muito nebuloso e controverso, como é o caso da natureza da inteligência. Que inteligência? A resposta não pode restringir tal capacidade a essa ou aquela espécie do reino animal ou a esse ou aquele sistema computacional, virtualmente habilitado a simular alguns processos cerebrais humanos considerados como "inteligentes". Por conseguinte, este trecho da pesquisa é excêntrico ao objetivo

geral, que logo a seguir será explicitado, e o leitor pode passar por cima dessas cerca de quinze páginas, sem prejuízo de entendimento do essencial do trabalho.

Maria E. Quilici Gonzalez [*apud* KHALFA(1995)] fez algumas reflexões sobre o assunto, remontando a épocas não muito distantes, para tentar capturar o entendimento do conceito de ***inteligência***, a começar pela ***corrente dualista*** e que considerava essa faculdade exclusiva dos seres humanos possuidores de uma alma (imortal), que lhes garantia o domínio da linguagem e das ações conscientes, não-mecânicas e responsáveis. As explicações para a inteligência deveriam ser buscadas, não no plano das causas que regem o mundo físico (material), mas sim, no plano das razões (abstratas) que os humanos poderiam ter para agir de forma a serem considerados inteligentes.

A perspectiva dualista sobre a natureza da inteligência e do seu domínio explicativo vem sofrendo profundas alterações com o fortalecimento da ciência. O germe do *behaviorismo*, bem ou mal, abriu caminhos para a investigação do comportamento inteligente, ao gerar hipóteses, muitas vezes convincentes, sobre a relação entre inteligência e aprendizagem associativa, segundo Maria E. Quilici Gonzalez [*apud* KHALFA(1995)].

Seguindo os rastros do *behaviorismo*, com outras nuanças, correntes dentro da atual Psicologia Cognitiva propõem um estudo da inteligência com base em processos automatizados (mecânicos) de resolução de problemas. Tais processos envolvem essencialmente a manipulação simbólica (IA cognitivista) ou mecanismos físicos de ajuste de pesos das conexões sinápticas de RNA, susceptíveis ao aprendizado associativo. Em ambas as linhas da IA, considera-se a inteligência (não sem polêmica) estendida às máquinas.

Apesar dos diferentes embates entre filósofos, lingüistas e psicólogos cognitivos (e seus aliados cientistas da computação), todas as perspectivas guardam algo em comum: existem vários tipos de inteligência, não facilmente comparáveis entre si.

Não é ilógico admitir a IA diante da compreensão de que há uma diversidade de processos inteligentes na natureza. Por esse prisma, SCHANK (1995) afirmou que o problema de se determinar a natureza da inteligência pode ser aproximado ao dos educadores, i.e., a inteligência é uma questão de aprendizado, de adquirir memória ou conhecimento básico de uma certa extensão e de desenvolver os mecanismos básicos de recuperação para tentar utilizá-lo. Como os pressupostos do autor são de que os problemas de IA têm características pedagógicas, se a IA possuísse uma extensão suficiente, para criá-la seria necessário um mecanismo cumulativo de aquisição.

Uma das críticas a SCHANK (1995) veio de Noam Chomsky, lingüista, criador da Gramática Gerativa-Transformacional, ao sustentar[58] que o aparelho genético inato que rege a linguagem é universal e exclusivo da espécie humana (CHOMSKY, 1998, p. 8-10).

SCHANK (1995) comentou a crítica do colega lingüista, dizendo que pelo seu paradigma a mente seria um apanhado de órgãos desconexos, especializados e voltados para a Matemática, Música ou outras aptidões. Assim, não existiria uma *explicação ambiental* das diferenças individuais no paradigma chomskiano.

Uma outra frente de críticas a SCHANK (1995) veio do filósofo John Searle, cuja essência, para não estender a descrição minuciosa do seu argumento[59], está na seguinte sentença: "Você aprende o que você vê". Para Searle, cães podem pensar; computadores não podem. SCHANK (1995) replicou, dizendo tratar-se de uma confusão que Searle fez com relação ao que é computador (*hardware*) e ao que é programa (*software*), entidades muito bem conhecidas pelos cientistas da computação, porque Searle insistia em identificar um *locus de compreensão* ou *locus causal* num único componente do sistema computacional: a UCP (unidade central de processamento), o que não corresponde à realidade, porque não se pode identificar num único item, seja ele de *hardware* ou de *software*, como sede da capacidade de se seguir regras num sistema computacional. Todavia, o argumento de Searle é considerado por alguns como definitivo contra a associação entre pensamento e computação, i.e., a relação entre o mental e o cerebral não seria análoga à relação entre *software* e *hardware* (ABRANTES, 1994).

Pelo que se viu, Chomsky e Searle têm algo em comum: adotam uma posição *essencialista* sobre a inteligência humana, com implicações parecidas no que tange à educação, porque, se Searle estiver correto, então a inteligência não pode ser aumentada. Nada poderá fazer algum ser já inteligente mais consciente do que já é, resultando que fornecer *regras*, *fatos* ou *experiências* para um sistema computacional[60] não os fará mais inteligentes, já que nenhum desses três itens têm algo a ver com a consciência (SCHANK, 1995).

A síntese da idéia de Roger Schank e de outros pesquisadores da controvertida linha da IA parece efluir da Teoria da Evolução de Charles Darwin. Como SCHANK (1995), SAGAN (1977), David Marr [*apud* RUELLE (1993)] e outros dessa linha crêem, não há muita diferença entre os processos básicos da inteligência humana e os de animais superiores.

---

[58] No original, R. Schank diz, na íntegra, que "esta sustentação vem por cima de impecáveis credenciais, politicamente corretas, que salvaram Chomsky da crítica implacável que outros receberam por defender as mesmas concepções".
[59] A experiência do quarto chinês de J. Searle, em ABRANTES (1994, p.12), (SHANCK, 1995) e (RUSSELL, 1995).
[60] Complexo formado de *hardware* e *software* bem diversificados (adaptação do pesquisador).

SCHANK (1995), além dessa crença, afirmou que a inteligência pode ser aumentada, ao contrário de Chomsky e Searle. Nesse ponto, é interessante citar alguns trechos do cientista David Marr, extraídos de sua obra *Vision* (1982). O autor idealizou projetar um robô inteligente pelo enfoque objetivo de um engenheiro e tirou várias conclusões desse exercício, entre elas:

" O problema da linguagem mostra que provavelmente não é fácil compreender a inteligência e que não é prudente limitar-se a uma única metodologia ... Juntando um instinto sexual, um sistema visual e alguns outros mecanismos do mesmo gênero, sem dúvida obteríamos um cérebro razoável para um rato ou um macaco. Mas o intelecto humano não será algo totalmente diferente, incomparavelmente superior? pois bem, talvez, não. Uma razão para se pensar que a diferença não seja extrema é que a diferenciação do cérebro humano levou relativamente pouco tempo na escala da evolução (alguns milhões de anos; e o desenvolvimento das linguagens complexas é, sem dúvida, mais recente) ... Noutras palavras, as aptidões especificamente humanas de utilização de instrumentos e de aprendizado de linguagens complexas deram-se provavelmente com uma relativa facilidade."

Marr ainda explicou que o cérebro humano e, conseqüentemente, a capacidade intelectual a ele associada, baseiam-se em mecanismos estritamente ligados ao problema da sobrevivência num certo tipo de ambiente e que a capacidade da linguagem e o sistema visual humano foram vantagens evolutivas de reforço à sobrevivência, apesar de admitir, nesse exercício de abstração, que o desenvolvimento do conhecimento científico, subproduto dos mecanismos superiores vantajosos adquiridos, é um mero acidente, visto como faltam ao cérebro humano certas funções básicas muito desejáveis para se fazer ciência: a aptidão para calcular de maneira rápida e confiável e a capacidade de memorizar grandes quantidades de dados, entre as de maior relevância.

SCHANK (1995) exemplificou vários casos para mostrar algum indício de que espécies animais superiores nem sempre agem por instinto, externando comportamentos que, se compreendidos, tornar-se-ão a pedra angular da inteligência humana. De todos os exemplos dados, o autor conclui que o comportamento racional segue o seguinte algoritmo: 1) Falha de expectativa; 2) Curiosidade; 3) Explicação derivada de uma experiência similar anterior (em virtude da memória); 4) Generalização; 5) Sucesso contingencial; 6) Repetição do ciclo, em razão de uma falha de expectativa num momento futuro.

SCHANK (1995), ainda preso mais a Chomsky do que a Searle, apontou algumas falhas das críticas dos dois autores à IA e lançou quatro questões sobre aspectos obscuros dessas críticas:

- Das cinco habilidades de um comportamento considerado como inteligente – *linguagem*, *memória*, *invenção*, *inferência* e *expectativa* -, todos constituindo apenas uma faceta da inteligência, Chomsky exclui considerações sobre a memória como sendo irrelevantes para o estudo da linguagem, tratando-a apenas como um mecanismo de retenção de planos estruturados de sentenças. A pergunta que se contrapõe a essa hipótese de Chomsky é a seguinte: "Pode-se falar do que não se sabe?" ou "Pode-se escrever sobre eventos dos quais não se pode lembrar ou imaginar?";
- Tanto Chomsky como Searle negam a hipótese da funcionalidade, tão explorada pela IA, fonte da maior contribuição dessa disciplina, que seria a criação de *modelos*[61] de processos das *aptidões mentais*. Esses modelos (tanto na IA como em campos afins) poderiam ser simulados em sistemas computacionais;
- A grande diferença entre a hipótese funcional (endossada pela IA) e a gerativa-transformacional de Chomsky é que a pedra angular desta última está na necessidade de explicar a capacidade intelectual de criar novas sentenças. Como diz CHOMSKY (1998): "Não é uma teoria, é um programa" (*programa minimalista*); pela perspectiva funcionalista, o que precisa ser explicado é a capacidade de criar novos pensamentos, ou de outra forma: "Como representar novas idéias em função de idéias antigas?" É um argumento simples: "O que torna alguém inteligente é o que ele sabe e o que é preciso para tornar computadores inteligentes é dotá-los de conhecimento";
- Chomsky não admite a existência de uma Psicologia que trabalhe com conteúdos mentais (crenças de senso-comum, propósitos, intenção, etc.); estes não poderiam ser modelados em computador porque não são científicos. Searle, ao contrário de Chomsky, claramente endossa a noção de que os conteúdos mentais são parte fundamental do conceito de inteligência, mas nega qualquer explicação computacional para tais conteúdos. O que diferencia ambos nas suas críticas à hipótese funcional é que Chomsky não considera os conteúdos mentais como fonte de problematização científica, com o que não concorda Searle. O que iguala ambos em termos de crítica é a tentativa de a IA trazer para o domínio computacional problemas de conteúdos mentais, i.e., *a IA é condenada por ambos por tentar fazer semântica*.

Quanto à segunda questão supracitada, vale ainda salientar que algumas linhas de pesquisa[62] da Psicologia e da Lingüística também utilizam a computação para seus propósitos, mas que são inadequados da perspectiva da IA, porque subutilizam a modelagem computacional, utilizando o sistema computacional apenas como máquina de calcular,

[61] Justamente o que se explora nesta tese.
[62] De fundo behaviorista.

putacional, utilizando o sistema computacional apenas como máquina de calcular, depósito de dados e traçador de gráficos, reamostrando o fenômeno mental de interesse por esses meios, pouco importando se os programas escritos são capazes de executar uma tarefa inteligente de qualquer tipo que seja. É justamente o contrário o interesse do cientista da computação: a modelagem da aptidão mental é a meta e não o fenômeno mental. O modelo computacional para aqueles lingüistas e psicólogos, razoável para as suas necessidades, diferencia um fenômeno do outro. Um modelo computacional para um pesquisador da IA serve para descobrir os óbices que se põem para tornar inteligente uma tarefa executada pelo sistema computacional.

Reforçando a quarta questão supracitada, MORA (1994) acentuou que a tese dos "universais sintático-fonológicos" dos transformacionistas, que explicariam as competências lingüísticas inatas de um sistema orgânico de comunicação, mereceu reparos dos seus próprios criadores. Eles orlaram, para isso, a tese *semântico-gerativa*, de natureza não-interpretativa, bem como o conceito do *signo invariante*. Daí emergiram mais fontes de reflexão para a Filosofia da Linguagem, na linha de contrapor as teses das tradicionais teorias semânticas (variação do significado) e os reparos transformacionistas.

Como já se viu (V. Figura 1.10) na concepção de um modelo para a IA, um AI só poderá existir se tiver um propósito e se puder conhecer para aprender. Um ambiente computacional capaz de ter um propósito, aprender e conhecer constituem as grandes barreiras que separam, de um lado, Roger Schank e os aliados da IA e da Psicologia Cognitiva, de outro, Noam Chomsky, lingüistas que endossam a Gramática Gerativa-Transformacional e psicólogos de tendências *behavioristas*; e ,ainda de outro, John Searle e seus aliados em idéias. Este é o cenário mais geral e as outras linhas que porventura existam não se afastam sobremodo dessas três examinadas.

SCHANK (1995) acreditava que há uma estreita ligação entre os problemas de Educação e os da IA e uma das questões que daí surgiu foi a seguinte: "Como os computadores poder ser empregados para melhorar a educação?" Para responder a esta questão e outras correlatas, o autor classificou a aprendizagem por sistemas computacionais em quatro arquiteturas:

- Arquitetura de ensino baseado em casos;
- Arquitetura de aprendizado incidental;
- Arquitetura de exploração direta de conexões de banco de dados orientados a vídeo;
- Arquitetura baseada na ação simulada.

A arquitetura baseada em casos foi considerada pelo autor como a mais promissora. Baseia-se em duas idéias centrais: 1ª) O especialista nesse sistema são arquivos de casos; 2ª) Bons professores devem ser bons contadores de histórias.

Portanto, de todos os enfoques da IA (V. Figura 1.11), o que mais interessa a SCHANK (1995) é o baseado na "equação": *IA = Fazer o computador aprender*. Segundo ela, inteligência implica aprendizagem; inteligência implica aprimorar-se ao longo do tempo. O problema com esse enfoque é que ninguém, até agora, realizou plenamente IA. "Onde está a IA?" é uma questão importante.

Para fazer IA é preciso um árduo trabalho de Engenharia de *Software*. Um programa de IA deve basear-se numa teoria que lhe ofereça o ensejo de ascender ou de evoluir. O problema de construir programas que exibam comportamento similar a um humano inteligente é que até agora não se chegou a um acordo sobre o que deve ser um comportamento inteligente. Para isso, é preciso construir programas (como os dos solucionadores de problemas, jogos de xadrez, etc.) que informem a respeito do comportamento humano, numa espiral ascendente de aprimoramento.

Na IA, o fato de um programa não ser capaz de melhorar não significa desqualificá-lo. As idéias contidas num programa podem ser idéias de IA, sem serem necessariamente idéias corretas de IA (SCHANK, 1995). Para resolver esse impasse, os programas adequados de IA devem ser capazes de processar uma imensa quantidade de informações.

Dessa forma, a IA se refere à representação do conhecimento. Mesmo um pequeno programa de computador[63], que fizesse algo desejado por alguém, poderia ser considerado como um programa de IA, caso se fundamentasse em idéias da IA (SCHANK, 1995).

A IA depende de sistemas computacionais que possuam conhecimento real. Isto significa que o *ponto crucial da IA reside na representação* desse *conhecimento*, na *classificação* por conteúdo desse conhecimento e na adaptação e *modificação* deste à luz da experiência e de seu uso; e o método de ensino baseado em casos é o que mais se aproxima desses requisitos (SCHANK, 1995).

PENROSE (1995) foi um autor que enveredou pela *inteligência matemática*, analisando-a em relação a outras formas de inteligência e entendimento humanos. Classificou esse tipo como forma extrema nesse rol, em virtude da natureza abstrata, impessoal e universal dos conceitos de que trata. Contudo, em seu artigo, demonstrou que o pensamento matemático incorpora outras qualidades também manifestadas pela capacidade humana de compreensão inteligente, tais como: intuição, bom-senso e apreciação da beleza.

Ao se questionar sobre o que vem a ser inteligência, o autor respondeu por um ponto de vista científico predominante, em que são os processos subjacentes cerebrais que controlam todas as funções corporais humanas por meio de cálculos extremamente complexos que, a princípio, poderiam ser traduzidos para um sistema computacional.

No decorrer dos testes, PENROSE (1995) concluiu que a Álgebra fornece meios de extrema utilidade para substituir intuições humanas por procedimentos de cálculo[64]. O uso efetivo da Álgebra requereria, com freqüência, uma boa dose de entendimento, sutileza e mesmo de talento artístico.

Aparentemente, PENROSE (1995) comungava com SCHANK (1995), mas ao longo do seu artigo não achou provável, pelo menos no atual estágio de desenvolvimento tecnológico, que os computadores pudessem simular o entendimento humano (incluído até o matemático, que para ele vai além de cálculos) pelo "cálculo cego" (cálculo sem entendimento).

Essa descrença de Roger Penrose talvez viesse da excessiva generalização com que o autor encarava a IA, pondo no mesmo nível os programas da IA cognitivista e as técnicas de RNAs e computação paralela dos conexionistas. Para ele, tudo seria "calculacional" e, portanto, reducionista.

Utilizando o Teorema de Kurt Gödel[65] (em operações aplicadas aos números naturais) como base de outros testes que descreveu em seu artigo, PENROSE (1995) demonstrou que a *inteligência matemática* não é de fundo meramente "calculacional". Após os testes, o autor concluiu que: "Os matemáticos não usam um procedimento calculacional comprovadamente correto para verificar a verdade matemática".

PENROSE (1995) acreditava que a capacidade de compreender foi um ganho seletivo para a espécie humana e que a compreensão da Matemática também foi um incidente vantajoso, não fugindo à regra do que, na sua essência, ocorre com outros mecanismos que não têm nada de tão abstrato, universal e impessoal, como p.ex.: a intuição.

Nesse ponto é interessante citar HUISMAN (1976). Ao tratar de Filosofia da Ciência, o autor comparou a intuição ao raciocínio. A *intuição* é um *conhecimento imediato* que é dado pelos órgãos dos sentidos, uma "visão" (*tueri*; significado primitivo do latim: *ver*). Conforme Kant: "A intuição é todo o conhecimento que se relaciona imediatamente com os objetos. Já o *raciocínio* não é imediato, mas procede *por mediação* para buscar ou gerar conhecimento.

---

[63] Caso do PROFAX.

[64] V. caso dos modelos matemáticos de avaliação de SS.

[65] "Se um sistema é demonstrável, ele é incompleto e se um sistema não é demonstrável, ele é completo" (1930). Sobre críticas às traduções desse teorema, V. TUFFANI (2002).

Enquanto a intuição revela ao homem realidades singulares e seres concretos, o raciocínio progride por meio de conceitos e de idéias gerais e abstratas. O raciocínio opera os conceitos por meio de silogismos, de hierarquias, exigindo, portanto, linguagem, que não pode existir sem palavras. Como o raciocínio se fundamenta na linguagem, pode ser comunicável, o que não ocorre com a intuição, inexprimível, porque se apóia numa realidade singular. As intuições humanas são experiências individuais, solitárias, inexprimíveis, incomunicáveis, ao passo que os homens só podem se entender por palavras, símbolos e raciocínios derivados destes signos.

HUISMAN (1976) continuou suas reflexões de cunho comparativo, dizendo que tanto a intuição como o raciocínio têm o seu lugar no conhecimento: a intuição fornece a matéria do conhecimento e o raciocínio se efetua com base nos dados fornecidos pela intuição.

E como HUISMAN (1976) e PENROSE (1995) se encontram nas idéias? Justamente no ponto em que o primeiro começou a conjecturar sobre o domínio do pensamento puramente abstrato, no domínio da Matemática, também asseverando que a intuição participa desse domínio, assim como participa dos mecanismos cognitivos que levam à descoberta.

HUISMAN (1976) analisou os *silogismos* e as *demonstrações* matemáticas. Os primeiros, modelos do rigor, foram classificados como "solenes futilidades", já que a conclusão já vem incorporada nas premissas e nada de inusitado acontece. Na demonstração ou nas deduções, entretanto, "às vezes se espera o resultado do cálculo com uma avassaladora ansiedade".

O processo de generalização, i.e., o progresso em relação a um conhecimento anterior, acompanhado da fecundidade da intuição adivinhadora (descoberta) são aspectos que PENROSE (1995) qualificou como irredutíveis a cálculos. Este autor acreditava que a natureza da inteligência só será compreendida, quando todos esses aspectos, para ele irredutíveis a formalismos lógico-matemáticos, puderem ser explicados.

GONZALEZ (1994), por outro lado, procurou estabelecer bases para uma Ciência Cognitiva independente, a fim de que seja cientificamente instituído um campo de estudos para explorar os problemas controvertidos sobre a natureza da inteligência.

A autora tinha por pressuposto que se o conceito nebuloso de informação não for devidamente explicitado, muitas das hipóteses (dos cognitivistas da IA, p.ex.) de utilizar os computadores para simular o comportamento humano inteligente ficam comprometidas.

A mesma autora asseverou em seu artigo que há muitas tentativas de explicitar o conceito de informação, mas que ainda não se chegou a uma definição de consenso. Ela distinguiu duas linhas de pensamento nessa busca de definição sobre informação: uma linha ad-

mite que a informação é um produto abstrato da mente consciente, noção que se afasta em demasia da outra linha (matéria e energia), cujos estudiosos consideram a informação como um produto objetivo (fato) e independente de qualquer ser consciente.

Diante dessas duas linhas de pensamento, restou a GONZALEZ (1994) optar por uma variante da segunda, proposta por J. J. Gibson, em 1966, na obra: "*The senses considered as perceptual systems*", para ter condições de partir de um marco teórico. Nessa obra, a informação é uma entidade que existe independentemente da existência de qualquer atividade consciente e interpretativa de um indivíduo e tal noção contribuiria bastante, segundo a autora, para trazer mais luz aos estudos das *representações mentais*[66], especificamente relacionadas às percepções visuais dos modelos conexionistas.

Essas representações mentais são caracterizadas por relações de analogia *neurônio-símile* (GONZALEZ, 1994) ou, como estabelece PASSOS (1990): neurônio biológico = elemento de processamento (EP) no modelo de RNA.

A justificativa de GONZALEZ (1994) para estudar os modelos conexionistas de representação mental pelo enfoque de J. J. Gibson, deveu-se à firme posição deste pesquisador de aliar-se à escola da percepção direta (*direct perception*) ou *eliminativista*, que é contra a tese da representação (*representacionalista*).

J. J. Gibson recusava-se a admitir a existência de representações mentais como entidades abstratas na percepção ou no comportamento humano. Mesmo assim, GONZALEZ (1994) acreditava que ele poderia ter nutrido interesse pela tese conexionista sobre representações mentais, como se verá adiante. Entretanto, o que mais chamava a atenção da autora, além desse suposto interesse de Gibson pelas representações mentais, era a plausibilidade de investigar o emprego do conceito de informação desenvolvido pelo pesquisador nos modelos conexionistas, conceito que não pressupunha qualquer intérprete consciente para manipulá-la. É um *constructo* de informação em bases mais objetivas e que também recebeu contribuição de trabalhos alinhados com o dela, na teoria e na metodologia, como os de C. Shannon e W. Weaver, F. Dretske, B. Küppers, K. Sayre e de T. Stonier.

O que mais sobressai do trabalho de GONZALEZ (1994) não está numa análise mais profunda sobre o que é inteligência, como se viu antes nos outros autores. Ela estabeleceu em seu trabalho, no meio científico nacional, um critério epistemológico distinto do que é até seguido por esta tese, que admite a IA como disciplina da Ciência da Computação, acatando as cautelas de ABRANTES (1994), que ainda não reconhece um campo científico indepen-

---

[66] V. "estados mentais" da tese cognitivista-funcionalista descrita no subitem 1.4.

dente como o da Ciência Cognitiva. GONZALEZ (1994), ao lado de alguns autores estrangeiros, como EYSENCK (1994), contrariamente, admitiu em seu trabalho a Ciência Cognitiva como uma área multidisciplinar de estudos e põe a IA e as RNAs como quase-disciplinas (vertentes de estudo ou linhas de pesquisa) dessa Ciência Cognitiva, declarando que há mais ou menos trinta anos o termo que denota esse campo científico já havia sido consagrado por E. Scheerer, em seu trabalho: "Towards a history of Cognitive Science" (1973).

Para sustentar as declarações de fundo epistemológico do parágrafo anterior, a mesma autora citou o objetivo central e até os pressupostos da Ciência Cognitiva:

- Objetivo central: elaboração de modelos e teorias científicas dos processos cognitivos humanos;
- 1º Pressuposto: o estudo do conhecimento humano requer uma investigação das capacidades de representação (*representacionais*) e computacionais da mente;
- 2º Pressuposto: o estudo científico da mente deve ser desenvolvido com base numa perspectiva interdisciplinar;
- 3º Pressuposto: os computadores, apesar de diferirem materialmente dos organismos, fornecem bons modelos para o estudo do sistema cognitivo humano.

Como se percebe, o terceiro pressuposto possui forte influência funcionalista e a autora definiu duas vertentes de estudo dentro "da sua" Ciência Cognitiva, oriundas desse paradigma, às quais denominou de: FLC (*funcionalismo lógico-computacional*) e FNC (*funcionalismo neurocomputacional*), que em nada diferem de outras já consagradas subdivisões disciplinares, em que a FLC seria comparada à linha cognitiva, *cognitivista* ou descendente da IA e a FNC, à linha conexionista, ascendente ou biológica da IA.

A novidade é que a autora não mais colocou os estudos conexionistas como inclusos na disciplina de IA. Ela denominou a vertente FLC como IA e criou uma genealogia diferente para o já consagrado campo conexionista de estudos, pondo-o como descendente-irmão da FLC e "pendurado na sua" Ciência Cognitiva: a FNC ou também *Conexionismo*, ou RNA ou Sistemas de Processamento de Informação em Paralelo. Essa nova disciplina é normalmente enquadrada no meio acadêmico como sendo uma linha de pesquisa da IA[67], ou como um conjunto de técnicas e idéias de ponta da linha de pesquisa de aprendizagem da IA, como consta em autores de tendência unificadora do gênero de RUSSELL (1995).

Apesar de não seguir a tipologia de GONZALEZ (1994), este pesquisador admite que as delimitações disciplinares ficam mais claras dessa forma; mas este é um problema fecun-

---

[67] Juntamente com as redes de crença, oriundas da Teoria Bayesiana.

do para a Filosofia da Ciência ou para a Teoria do Conhecimento (V. glossário) e não será tratado em mais profundidade, mesmo nesta seção excêntrica da tese. Para ilustrar esta parte do assunto, apresenta-se a Figura 1.14 como síntese das idéias de GONZALEZ (1994).

A Figura 1.14 não incluiu algumas ciências citadas por GONZALEZ (1994), que também interferem na constituição da Ciência Cognitiva, como a Neurofisiologia e a Física, mais preocupadas com os elementos materiais, segundo a tese da autora. A omissão teve o fito de não congestionar a figura com linhas e texto, tumultuando o entendimento principal que se pretende, que é o de mostrar o afastamento da FNC em relação à FLC, pelo menos dentro da linha de pesquisa encetada por GONZALEZ (1994).

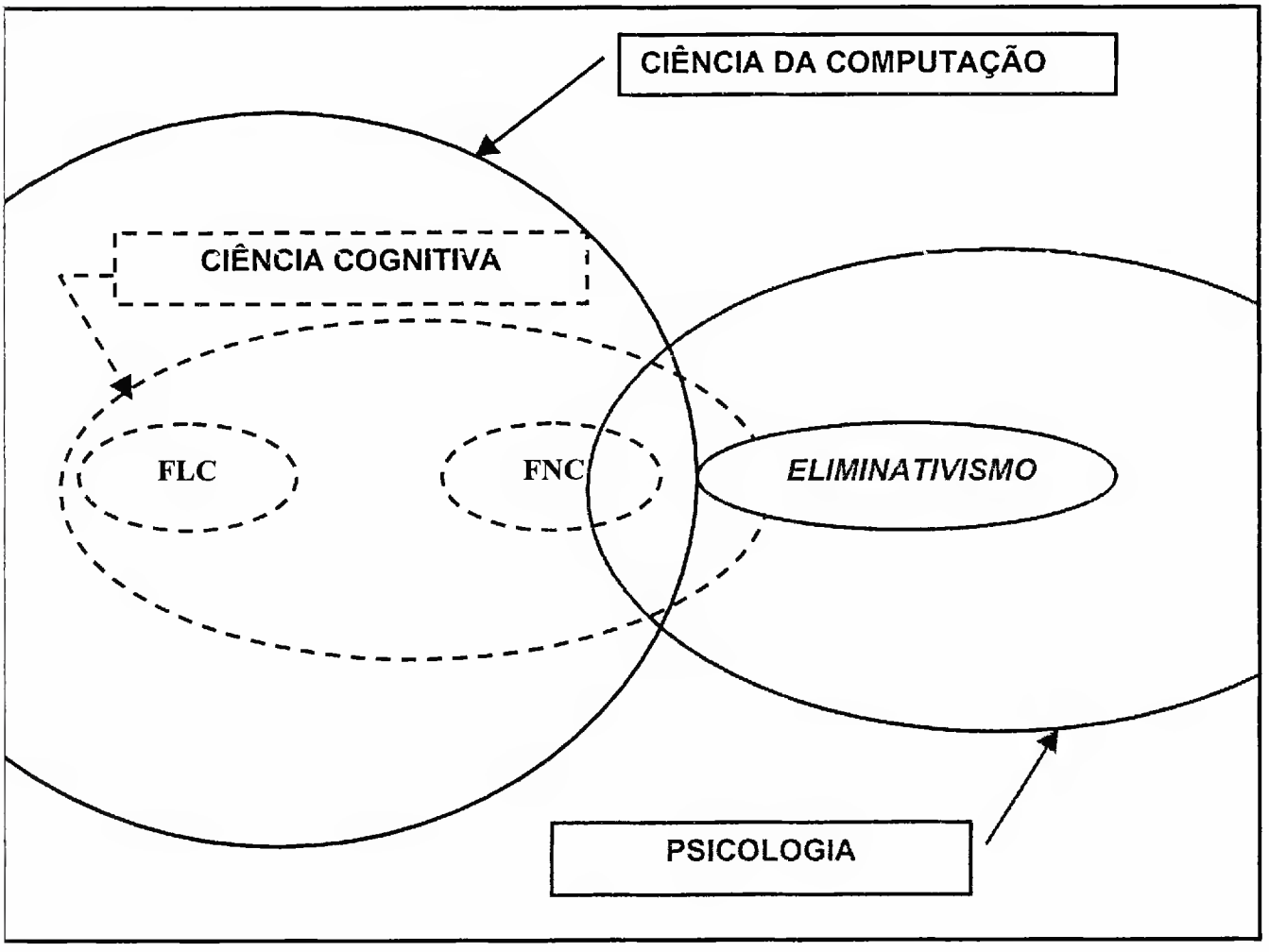

Figura 1.14: A Ciência Cognitiva e suas vertentes.

A grande contribuição de GONZALEZ (1994) parece ser, além da proposta de taxinomia para as ciências cognitivas, o preenchimento de lacunas conceituais entre as vertentes

cognitiva e conexionista da IA, caracterizando cada uma como áreas independentes de conhecimento. Os *constructos* que operariam esta diferenciação seriam o de *informação* e o de *representação mental*[68], de certa forma trazendo a área conexionista para mais perto da área de influência *gibsoniana* ou *eliminativista*, como a Figura 1.14 representa de forma gráfica.

Apenas para relatar muito rapidamente o *conceito de informação* de Gibson, já que esta tese está sob influência de um curso de CI, é instrutivo passar os olhos as idéias deste autor. Para ele, informação é um conceito de caráter dual: por um lado, conta com um elemento objetivo que existe no meio-ambiente, em decorrência de suas propriedades relativamente permanentes (os *invariantes estruturais* e *transformacionais*), mesmo que ocorram sensíveis mudanças ao seu redor. Esses invariantes são denominados por Gibson como *informação sobre algo*. A segunda face da informação é conhecida como *informação para alguém*. De caráter relacional, este aspecto da informação diz respeito àqueles padrões que o organismo está preparado para captar, dado o seu estado evolutivo[69].

Esta tese tem seu marco teórico nas teorias da linha denominada por GONZALEZ (1994) de FLC, sobre a qual não recaiu o interesse dessa autora. Do lado da FLC há muitas restrições contra a FNC, especialmente por tratar-se de um campo que depende de muitos sucessos no sem-número de algoritmos que suas técnicas utilizam, a fim de que seus pesquisadores conquistem a necessária consagração científica do campo. Certas aplicações como reconhecimento de voz, reconhecimento de escrita corrida e auxílio na condução de veículos, todas relatadas por RUSSELL (1995, p. 584-587), já se mostraram capazes de ser incorporadas pelo mercado.

O que tem tornado a FNC muito atraente para os cientistas são as pesquisas na área da aprendizagem e a regular utilização da computação paralela e mesmo a analógica, ideal para simular com mais realismo certos fenômenos físicos ligados aos processos de ligação sináptica e de equilíbrio de uma RNA. Sendo assim, é preciso esperar por mais testes e trabalhos teóricos nesse campo.

Já foi mencionado que outros autores (nacionais e estrangeiros) já adotaram, como GONZALEZ (1994), a denominação de Ciência Cognitiva como um campo independente do conhecimento, porém nem todos o fazem sem críticas e, entre eles, está a SEP[70] (2001), que reúne uma lista das maiores críticas que se fazem à Ciência Cognitiva, todas mais ou

---

[68] V. *esquema* e *representação* ou imagem *mental*, na p. 146.
[69] Para maiores particularidades, V. GONZALEZ (1994, p.140-145).
[70] ***Stanford Encyclopedia of Philosophy.***

menos centradas na alegação dos pesquisadores desta área de que "a mente humana funciona na base da representação e da computação". E não adianta pertencer à linha cognitivista, admitindo a analogia: mente = proposições lógicas + conceitos + regras + analogias + dedução + busca + comparação + recuperação, ou à linha conexionista, mudando a analogia de mente para cérebro e admitindo que as ligações sinápticas dos neurônios possam ser simuladas por algoritmos que disparam processos de cálculo se espalhando por uma estrutura de rede neural artificial. As críticas aos dois modelos são mais desafios e tais desafios são produtivos de duas formas: servem de fonte de trabalho reflexivo para a Filosofia e de motivação para os cientistas cognitivos trabalharem, tendo em mira eliminar as críticas ao sobrepujar os desafios. Essas críticas (desafios) são as seguintes:

- Emoção: os cientistas cognitivos negam o importante papel das emoções no pensamento humano;
- Consciência: os cientistas cognitivos ignoram a importância da consciência para o pensamento humano;
- Mundo: os cientistas cognitivos desconsideram o significativo papel do meio físico no pensamento humano;
- Sociedade: o pensamento humano está fortemente inter-relacionado com o meio social e a Ciência Cognitiva não trata desse aspecto;
- Sistemas dinâmicos: a mente é um sistema dinâmico (no sentido mais amplo do termo) e não automatizado, como um computador, repleto de restrições;
- Matemática: os resultados matemáticos mostraram que a atividade intelectual humana não pode ser modelada pelos padrões computacionais; o cérebro não trabalha exatamente[71] dessa forma; provavelmente o cérebro seria mais um computador quântico.

A Filosofia da Ciência propõe as seguintes questões de ordem metodológica para esse campo científico que emerge:

- Qual é a natureza da representação?
- Que papel os modelos computacionais representam no desenvolvimento das teorias da Ciência Cognitiva?
- Qual é a relação que existe entre os aparentemente concorrentes argumentos ligados à mente, quando se trata de processamento simbólico (cognitivismo), redes neurais (conexionismo) e sistemas dinâmicos?

---

[71] Exceto em alguns poucos casos, pode haver uma certa semelhança (nota do pesquisador desta tese).

Encerrando esta seção reflexiva da tese, nada mais auspicioso que coroar o teor do título deste subitem com uma visão geral do trabalho de CRUZ (1994), que se interessou pela IA em virtude das investigações teóricas que realizou com outros pesquisadores sobre as limitações do uso de modelos formais em Biologia. Os resultados de sua análise convenceram-no da necessidade de associar métodos da IA a modelos matemáticos para explorar a semântica de expressões biológicas, para avaliar a ambigüidade dos conceitos envolvidos (o de inteligência, p.ex.) e para propor especificações de sistema que viessem a minimizar as inconsistências das previsões desses modelos em face das observações factuais.

A questão preliminar e angustiante que surgiu para CRUZ (1994) e seu grupo foi a seguinte: "Em que medida o computador é inteligente?" O autor se viu, então, diante de um complexo problema filosófico, no momento da concepção do seu problema de pesquisa: como lidar com um sistema computacional dotado de IA, que deve processar informação, se não se pode definir o que seja informação com precisão? Diante disso, o grupo de pesquisa optou por abandonar momentaneamente as questões de ordem filosófica; explicitar uma tendência bem aberta para definir o conceito de inteligência e levar em consideração o caráter evolutivo da inteligência[72] (na verdade, são três objetivos de pesquisa).

Sobre a última consideração de CRUZ (1994), no parágrafo anterior, uma importante diretriz de trabalho foi estabelecida pelo o autor, assim como uma interessante perspectiva se esboçou para o presente trabalho de pesquisa sobre SS, que pode ser expressa pela seguinte sentença: Não se pode negligenciar a inteligência como um processo biológico, que se expressa diferentemente através de toda a escala filogenética, i.e., a inteligência é um processo evolutivo [como pensava SCHANK (1995)].

Voltando aos três objetivos de pesquisa de CRUZ (1994), tendo em consideração o 2º objetivo (inteligência = conceito aberto) e o 3º objetivo (definição operacional de inteligência pelo modelo de Guilford, de base evolutiva), pode-se esquematizar uma definição operacional de inteligência[73], atribuindo-lhe 120 habilidades[74], distribuídas em cinco categorias de operações, quatro categorias de conteúdos e seis categorias de produtos. São as combinações (produtos polinomiais em que os fatores são as categorias) entre operações, conteúdos e produtos, tendo como controle um quadro bem amplo de experimentos empíricos realizados em grupos de indivíduos, que produzem as habilidades de Guilford. A Figura 1.15 dá uma idéia desse modelo de Guilford, relatado por CRUZ (1994).

---

[72] A obra de referência para o trabalho de CRUZ (1994) foi "The nature of human intelligence", de J. P. Guilford (1967).
[73] A única mais palpável desta revisão de literatura.
[74] Segundo CRUZ (1994), em 1984, em conseqüência da obtenção de novos dados empíricos, as habilidades foram aumentadas de 120 para 150.

Para complementar CRUZ (1994), CILIATO (2000) descreve as 15 categorias das quais se derivam as 120 (5 x 4 x 6) habilidades de Guilford[75]. Em primeiro lugar, as categorias de operações (processos psicológicos básicos ou operações cognitivas):

- *Avaliação*: categoria que envolve as habilidades atinentes à tomada de decisão diante de um problema, sem hesitações persistentes, assim como a habilidade de avaliar a qualidade da decisão tomada;

Figura 1.15: Esquema do modelo de Guilford para a definição de inteligência.

- *Produção convergente* (pensamento convergente): categoria que integra um agrupamento de idéias divergentes num conceito unificador;

---

[75] Entre parênteses, a nomenclatura de CILIATO (2000).

- *Produção divergente* (pensamento divergente): categoria que se refere à facilidade de um indivíduo produzir uma variedade de hipóteses ou suposições perante situações problemáticas;
- *Memória*: categoria caracterizada pela capacidade de reter informações;
- *Cognição* (reconhecimento): categoria que envolve sensibilidade aos aspectos do meio-ambiente, consciência das alterações ocorridas nos estímulos externos e habilidade de dominar adequadamente o meio-ambiente;

Em segundo lugar, as categorias materiais ou de conteúdo:

- *Figural* (de figuras): imagens, desenhos, esboços, formas geométricas;
- *Simbólica* (símbolos): letras e numerais;
- *Semântica* (semântica): palavras e sentenças;
- *Comportamental* (comportamento).

Em terceiro lugar, as categorias de produtos cognitivos (informação processada):

- *Unidade*: uma única palavra ou idéia;
- *Classe*: conceito que representa um conjunto de unidades;
- *Relação*: interdependência, conexão ou correspondência entre unidades ou classes;
- *Sistema*: seqüência ou rede organizada de idéias ou conceitos;
- *Transformação*: mudanças ou redefinições de unidades ou classes;
- *Implicação*: análise das conseqüências.

Guilford afirmou que cada indivíduo é um composto particular de um grande número de habilidades intelectuais, sendo que cada uma delas é o resultado da combinação de três categorias de cada agrupamento enumerado (CILIATO, 2000).

Mais do que esmiuçar a descrição de cada categoria, o que mais importa entender é que: 1) Esse modelo fornece a definição operacional de inteligência; 2) CRUZ (1994) fez sua escolha epistemológica pela vertente cognitivista (lógica) e 3) O autor justificou não haver incompatibilidade entre sua escolha e a teoria de fundo evolutivo de Guilford[76].

CRUZ (1994) expôs o seu plano experimental exploratório para apurar uma definição para o termo *inteligência*.

Nesse ponto, não é necessário pormenorizar o plano de CRUZ (1994), já que os elementos conceituais (decomposição por traços semânticos[77], fixação de significado denotativo e conotativo, etc.) de que se utilizou serão analisados com mais profundidade na revisão de literatura e na metodologia.

---

[76] Denominada por CRUZ (1994) de *cognitivo-ambiental*.

[77] Fonte teórica: "The structure of a semantic theory", de J. J. Katz e J. A. Fodor (1963).

Por enquanto, importa apenas assinalar que, de posse desses elementos, CRUZ (1994) conseguiu delinear a ***distância semântica*** entre o sentido denotativo (V. Figura 3.7) e o conotativo de inteligência, que é a resposta para a questão inicial: "Em que medida a máquina lógica é inteligente?", resultado empírico que oferece ao pesquisador uma relativa (em hipótese alguma absoluta) idéia de quanto uma certa inteligência (animal ou não) está afastada em termos semânticos da inteligência de referência, que é a humana.

As contribuições do trabalho do grupo de CRUZ (1994) foram as seguintes:

- Abandonou a categórica pergunta: "Os computadores (máquinas lógicas) são inteligentes?", que exigiria um simples, absoluto e arriscado "sim" ou "não", pela pergunta: "Em que medida um computador é inteligente?", que estabelece uma distância entre a inteligência humana e a de qualquer outra entidade (animal ou máquina);
- Estimou valores (0: menos similar; 1: mais similar) de comparação entre sentidos de termos (no caso: *inteligência de máquina x inteligência humana*), ao utilizar a Teoria dos Conjuntos e a Teoria Semântica aplicada em componentes semânticos dos termos (os traços semânticos).

De certa forma, é sobre esta distância (primeira contribuição) que todas as metodologias de avaliação de SS (na revisão de literatura) tratam. CRUZ (1994), por conseguinte, arranjou uma solução lingüística e evolutiva para o problema.

A junção de uma teoria semântica com o formalismo de uma teoria matemática facilita metodologicamente tarefas de comparação (segunda contribuição). Essa é outra característica comum que será observada nas obras que serviram de referência para o método de construção do PROFAX e do PRONTO©, ora combinando teorias de fundamentação lingüística com a Teoria dos Conjuntos, ora com a Teoria da Geometria Analítica no Espaço Euclidiano, ora com a Probabilidade e Estatística.

2. O PROBLEMA E O OBJETIVO DE PESQUISA – ASPECTOS GERAIS

"Se você conhece o inimigo e conhece a si mesmo, não precisa temer o resultado de cem batalhas. Se você se conhece, mas não conhece o inimigo, para cada vitória sofrerá uma derrota. Se você não conhece o inimigo nem a si mesmo, perderá todas as batalhas." **(Sun Tzu – A Arte da Guerra – séc. V A.C.)**

2.1. Motivação e justificativa para a pesquisa

A origem do interesse pelo presente tema de pesquisa vem de cerca de quinze anos atrás, quando apenas a graduação deste pesquisador em Engenharia Cartográfica era o único elo entre o emprego das tecnologias da informação deste campo do conhecimento e a vocação latente pela investigação da natureza da informação geográfica.

O trabalho de engenheiro na linha de produção de documentos cartográficos não deixa muito tempo para reflexões sobre as causas de certas inconsistências no processo de comunicação homem-máquina e sobre outros fatores de natureza mais sutil, que podem potencializar diminutos "gargalos de produtividade" para o futuro e, pior ainda, acumular insatisfações no repertório de requisitos cada vez mais refinado de exigentes usuários.

A vocação latente pela pesquisa, num primeiro estágio, transformou-se em inquietação, quando eram adotadas soluções expeditas e de pouco alcance, sempre calcadas na mudança de *hardware*, em favor de mais desempenho, ou na mudança de *software*, quando o de emprego corrente não acompanhava mais as novas necessidades de produção[78].

O ambiente caótico desse período era uma constante no mundo de produtores e usuários de informação geográfica. E isto foi normal, porque era a transição da Cartografia convencional, de características artesanais, para a Cartografia Apoiada por Computador (CAC), que apesar de automatizar certos processos manuais da anterior, ainda era (e continua sendo) muito empírica e baseada no binômio *ensaio* e *erro*.

Esse período coincidiu com profundas mudanças nas TIs: os computadores deixavam de ser meras calculadoras e repositórios de dados para se tornarem valiosos agentes de auxílio aos projetos de engenharia (a era CAD ou *Computer Aided Design*).

---

[78] Quais seriam as bases dessas necessidades? Seriam elas objetivamente levantadas? Em grande parte dos casos, a resposta era: "Não!"

A pergunta que já se fazia àquela época e que veio a ressoar no XVIII Congresso Internacional de Fotogrametria[79] e Detecção Remota (CIFDR - Viena, 1996), baseava-se no seguinte: ***"É possível haver algum tipo de tratamento de dados cartográficos em meio digital que siga uma orientação mais próxima da finalidade da Cartografia, que é de natureza comunicativa?"***

O trabalho de BÄHR (1996), apresentado no XVIII CIFDR, traduziu o espírito da pergunta anterior e, no âmbito desta pesquisa, pôs termo à fase de inquietação. Foi a motivação para iniciar a busca por uma resposta, ou, pelo menos, de buscar uma coleção de alternativas que não se baseassem simplesmente em fundamentos da Lógica ou da Geometria (Matemática), no intuito de modelar um sistema que ajudasse a interpretar o fenômeno geográfico de uma forma mais próxima da linguagem humana.

Há outros caminhos não tão "ortodoxos" como os prognosticados por BÄHR (1996) e já trilhados por RODRÍGUEZ (2000), mas nem por isso menos científicos. Sobre estes últimos, alguns críticos podem criar polêmicas (até justas) sobre as lacunas que certos enfoques ainda suscitam, mas se estas polêmicas ainda não foram mediadas pela consagração da comunidade científica, abrem um sem-número de oportunidades de avançar as fronteiras de outros campos científicos, especialmente o das ciências cognitivas (Lingüística, Ciência da Computação, Neurofisiologia e Psicologia Cognitiva).

A principal ***motivação*** deste trabalho surgiu do estímulo negativo representado pela dificuldade de trabalho já circunstanciada e da decorrente expectativa positiva advinda do conhecimento sobre as iniciativas para abrandar ou eliminar muitas das dificuldades que ainda dominam o ambiente nacional de produção de dados cartográficos digitais. Essas iniciativas estão focadas na busca por requisitos de interoperabilidade entre SIGs utilizados pelas diversas organizações brasileiras (governamentais ou privadas).

Desses fatos é que surgiu o interesse, nesta pesquisa, pela descoberta de novos mecanismos de recuperação de informação, não mais baseados em simples comparação de cadeias de caracteres e tratamento estatístico; assim como na descoberta de novos mecanismos de integração de informações, mais comprometidos com a carga de bases de conhecimento e com a criação de funções que formalizem a similaridade semântica, que simplifica muito mais as tarefas de comparação de entidades representadas em níveis mais baixos, como em bases de dados, dispensando a interação do usuário não especializado com os bastidores dessas TIs.

---

[79] V. glossário.

Para **justificar** este trabalho é preciso citar duas razões, que podem ser denominadas de: 1ª) Razão interna e 2ª) Razão externa.

A *razão interna* funda-se no desconhecimento dos produtores e usuários brasileiros de informação geográfica (IG). Do lado dos produtores, o problema não está ligado ao interesse dos técnicos e pesquisadores pela busca de outras alternativas de mapeamento, mas ao pouco fomento governamental e privado na capacitação desses profissionais, o que os marginaliza em relação ao estado atual do conhecimento no campo da CIGeo no mundo. Do lado dos usuários, o problema é mais grave, já que é também a causa indireta da falta de recursos para os produtores. Para sintetizar esta visão, é preciso dividir os usuários em dois grupos: os que decidem sobre a aplicação das tecnologias de *geoprocessamento* e os que as utilizam. O traço comum para ambos os grupos, com agravante para o primeiro, também é o desconhecimento sobre o uso e o potencial dessas tecnologias.

Diante desse quadro, tem havido um incansável[80] esforço da comunidade *geocientífica* nacional para cobrir a lacuna de conhecimento apontada. São seminários, simpósios e congressos, em que a tônica tem sido diminuir a carga sobre assuntos de complexidade científico-tecnológica e procurar focar as matérias e os debates sobre o uso de cada tecnologia, trazendo usuários refratários para a órbita do *geoprocessamento*.

O que vem por trás de todo esse cenário de capacitação insuficiente e falta de visão estratégica na produção e uso da IG é um problema ainda maior (na verdadeira acepção de "maior"): o ***vazio cartográfico*** do Brasil e as dificuldades de manter atualizados os documentos cartográficos que se produzem pelas instituições governamentais responsáveis pelo mapeamento.

Para se ter uma vaga idéia sobre esse problema do *vazio* cartográfico, que também justifica esta pesquisa, basta dizer que os documentos cartográficos (cartas topográficas na escala de 1: 25.000, i.e., 1 cm na escala da carta = 25 m no terreno) para cobrir a parte do país em que são necessários projetos urgentes de engenharia (construção de estradas, pontes, barragens, etc.) não chega a 20%; e esta cifra precária fica ainda mais comprometida com o fato de que um documento cartográfico, por mais que tenha sido agilizada a sua produção por processos automatizados (CAC), a rigor, já sai da linha de produção desatualizado! É um jogo de forças em que, se não forem aplicadas novas idéias, a Cartografia nacional poderá mergulhar numa espiral descendente de atendimento aos anseios básicos da socie-

[80] O pesquisador não tem informações atualizadas se o esforço tem produzido os efeitos desejados, como, p.ex., os iniciados no CEPAD/CONCAR, em 1997 (V. glossário).

dade nas políticas de uso e ocupação do solo, saneamento, saúde, agricultura, enfim, em todas as políticas em que o pano-de-fundo geográfico entra como suporte da política.

A *razão externa* funda-se na explosão de alternativas de mapeamento que surgem no exterior. Muitas estão atreladas ao vetor tecnológico, promovendo comodidade ao usuário e rapidez no tratamento da IG, desde a fase da coleta de dados até a reprodução final dos resultados. Tais alternativas estão calcadas no aprimoramento de elementos específicos de *hardware* e em *softwares* mais apropriados para automatizar processos convencionais de mapeamento. Outras alternativas, no entanto, fogem aos padrões convencionais e, apesar de ainda não terem alcançado um nível de adequação[81] para serem comercializadas, seus protótipos e projetos-pilotos mostram-se extremamente promissores.

Para resumir o que significam essas alternativas não-convencionais, ainda longínquas do cenário de pesquisa brasileiro e que também justificam esta pesquisa, é bom citar algumas idéias de pesquisadores que estão na vanguarda dessas alternativas: Max J. Egenhofer e Werner Kuhn, ambos docentes oriundos de universidades alemãs e, atualmente, lecionando na Universidade do Maine (EUA) ou dirigindo projetos de pesquisa financiados pelo NGCIA (Centro Nacional de Análise de Informação Espacial dos EUA).

As idéias básicas sobre essas alternativas não-convencionais de mapeamento já foram enfocadas na introdução, tendo "interação" por palavra-chave. Esta interação começou na década de 70, com a comunicação homem-máquina dominada pelo estilo *das linhas de comandos* em sintaxe de linguagens estruturadas de busca (SQL), que orientavam as pesquisas dos usuários em bases de dados geográficas. Depois, o avanço de deixar escondida esta sintaxe por trás de *janelas com ícones e menus*[82] *de ferramentas embutidos em botões*, artifícios engenhosos que engrossaram as fileiras de usuários dessas TIs, porque o esforço de aprendizagem era diminuído, não se devendo mais apelar para a memorização que era necessária no estilo de linha de comando.

Na década de 90, o estado-da-arte para os usuários de interfaces gráficas de SIG fundava-se no *padrão WIMP* (*windows*, *icons*, *menus*, *pointing*), que adicionava algumas inovações ao estilo anterior, como implementação de pacotes de certas funções muito utilizadas (cálculo de distância, área, volume, etc.) noutros ícones e algumas funcionalidades de arrasto e indicação com *mouses* e cursores de CAD.

Até então, o paradigma prevalecente de apresentação visual e de interação homem-máquina era o de realizar uma consulta numa BD geográfica e receber os seus resultados

[81] Menos pela confiança e mais pelos efeitos que a quebra de paradigma produziria no parque tecnológico já instalado.
[82] Menus *pull-down*.

em mapas e tabelas, processo que era ainda um tanto dependente das capacidades dos usuários em lidar com algumas particularidades dos SIGs. Com o surgimento das bibliotecas digitais e a distribuição de dados pelo serviço www da Internet, os usuários passaram a ter acesso a uma quantidade ainda maior de informação para consulta. A correspondente vertente de recuperação de informação (IG) na área da Cartografia foi materializada pelas mapotecas digitais.

Embora grandes avanços tivessem ocorrido em relação às técnicas de acesso à IG, as aplicações de *geoprocessamento* ainda exigiam dos usuários certa especialização. A interface gráfica com o usuário ainda era centrada na resolução de problemas espaciais e no apoio à decisão que dependesse dessa área, como por exemplo: seleção de temas (*layers*) numa base de dados, identificação de objetos espaciais por coordenadas, toponímia (nomes), estabelecimento de relações espaciais elementares (criação de figuras complexas com base em figuras mais simples, p.ex.) e modificação de parâmetros (cor, peso e estilo das linhas, etc.) de exibição ou de impressão de documentos cartográficos. Isto ainda era (e ainda é insuficiente). É preciso dinamizar ainda mais os recursos de visualização para reduzir a complexidade de um SIG e isto implica mais investimento.

Na ausência de um modelo conceitual de interface gráfica, soluções improvisadas, mais orientadas a certas necessidades específicas, foram implementadas, acompanhando a tendência geral das casas de *software*, preocupadas em comercializar soluções imediatas para sistemas interativos, bastando notar a profusão de processadores populares de texto.

Com a IG, todavia, o problema é mais complexo que o puro texto, como já tanto se explicou, não obstante também terem surgido sistemas muito personalizados, como p.ex. as *cartas eletrônicas* ou *sistemas de auxílio à navegação* ou à condução de veículos (*car navigation systems*).

A partir daí intensificaram-se as pesquisas sobre formas mais naturais de comunicação homem-máquina, que foram além da apresentação visual e começaram a explorar outras formas sensoriais de entrada e saída num SIG: o ***esboço*** (*sketching*) e a ***gesticulação*** (*gesturing*), como inovações no campo da visualização, entrada por dispositivos de reconhecimento de voz, exploração da audição e até mesmo pesquisa sobre dispositivos tácteis de entrada e saída.

Com o crescente interesse em recursos de multimídia, agregados à rede mundial, tornaram-se quase ilimitadas as necessidades e possibilidades de representação espacial da informação: manipulação de fenômenos por dispositivos holográficos, movimento, simulação, enfim, possibilidades outrora "escondidas" em bases de dados e exploradas por aplicativos

difíceis de manusear por um indivíduo comum estão, agora, ao alcance de uma parcela significativa de usuários a preços não muito proibitivos. Parece muito apropriada a expressão com que Werner Kuhn e Max Egenhofer qualificaram essas formas de comunicação: ***"Intergalactic data speak[83]"***.

É bom que se diga que todos esses desenvolvimentos se deram fora da área de aplicações em *geoprocessamento*. Foram todas contribuições interdisciplinares entre as pesquisas em SIG e teorias cognitivas (áreas de interação, percepção e colaboração) que serão vistas na exposição da natureza do problema.

2.2. Natureza e formulação do problema geral de pesquisa

Uma inquietação comum e mais ou menos simultânea sacudiu os ambientes *geocientíficos* do planeta em meados da década de 90. As causas dessa inquietação já foram razoavelmente elucidadas, mas a natureza do problema primordial que eclodiu nessa época ainda não foi detidamente analisada, porque transcende o domínio dos *bits* e *bytes* e penetra numa zona ainda debilmente conhecida pelo homem: a sua própria mente.

2.2.1. Considerações gerais

Mesmo sem ainda penetrar na revisão de literatura (Capítulo 3), já se pode dizer que as linhas de pesquisa sobre interface-gráfica e comunicação homem-máquina[84] estão na pauta da Ciência Cognitiva, Sociologia e Economia (EGENHOFER, 2001a). A primeira estuda como as pessoas pensam, se comunicam e formulam seus problemas sobre informação geográfica e como isso afeta o uso do *software ad-hoc*. Os estudos sociológicos investigam o papel das TIs na sociedade, especialmente pelo enfoque do uso dessas tecnologias. Já a Economia enfoca técnicas de reengenharia e de engenharia reversa (V. glossário) a serem exploradas pelas organizações e de que forma essas organizações podem compartilhar bases de dados.

Essas são, portanto, as demandas multidisciplinares que estão surgindo e o desenvolvimento de SIGs não pode mais ficar à margem delas, restrito apenas à parte mais simples das aplicações de natureza geográfica, aquelas ainda atreladas às soluções matemáticas, sem alcance qualitativo. Tais soluções, via de regra, são implementadas por produtos de

---

[83] EGENHOFER (2001a)

[84] Também chamada de interação homem-máquina por PRESSMAN (1995, p. 195); HCI – *human-machine interaction*.

software que exigem habilidades especiais dos usuários para manipulá-los e para compreender o que os resultados limitados expressam sobre o fato original (geográfico).

A natureza do problema que será apresentado a seguir está no campo cognitivo. Trata-se de avaliar julgamentos humanos sobre semelhanças, envolvendo relações que podem ser estudadas aproximadamente pelo enfoque matemático (relações de equivalência[85]), porém sem o seu rigor formal, compensando tal rigor com observações empíricas, cuja metodologia, em grande parte, veio da Psicologia Cognitiva, por intermédio da IA.

Conforme RODRÍGUEZ (2000, p.1), *grosso modo*, avaliar ou julgar a similaridade entre "coisas" (reais ou abstratas) é um processo cognitivo que implica decompor essas "coisas" em elementos que se pareçam ou que difiram entre si; e os seres humanos exercitam essa habilidade razoavelmente bem.

Ainda segundo essa autora, o tipo de julgamento desenvolvido no dia-a-dia dos indivíduos, relacionado à avaliação de semelhanças e diferenças entre estímulos percebidos do mundo, é intuitivo, subjetivo e não demonstra qualquer característica subjacente de natureza formal ou matemática.

Nos sistemas de informação (SIs), a avaliação desse tipo de relação de equivalência – a similaridade -, desempenha um importante papel na ***recuperação*** e na ***integração*** de informações, assim como na manutenção dos dados.

Como um SIG é um SI especial, em que os usuários de informação geográfica (IG) possuem diferentes níveis de conhecimento e como é uma tarefa complexa definir (documentar) precisamente as entidades espaciais que podem ser representadas nesses sistemas, um critério de similaridade num nível bem alto de abstração, calcado em fundamentos formais mas não necessariamente normalizado, pode vir a ser útil na ingente missão de integrar[86] gigantescas bases de dados espalhadas pelo planeta, construídas sob os mais diversos enfoques de grupos de usuários, que vasculham por informações nessas bases pelas ainda muito utilizadas linguagens estruturadas de consulta (SQL), com resultados não muito exatos em relação às reais intenções de busca dos usuários.

Esta tese está na mesma linha de pesquisa da de RODRÍGUEZ (2000), preocupada com o aspecto semântico das entidades espaciais, propondo-se, numa primeira fase, um modelo mais simples que o da autora para avaliar a similaridade semântica entre essas entidades, para, a seguir, utilizar o modelo da autora consoante as particularidades do estudo-de-caso levantado pelo problema geral e esmiuçado pelos três objetivos específicos.

---

[85] Na revisão de literatura, será definida a SS, verificando-se que ela é um tipo especial de relação de equivalência.
[86] Tornar interoperáveis os aplicativos e sistemas dedicados a essas bases.

As primeiras pesquisas nesta área[87] começaram em meados da década de 90 e contemplaram mais os aspectos geométricos. Como este trabalho não se ocupa com as propriedades geométricas das entidades espaciais, seu foco se aplica sobre as propriedades de ordem cognitiva que a avaliação da similaridade semântica suscita em relação ao domínio espacial, deixando para um trabalho futuro a tarefa de integração entre esses dois ramos.

O termo entidades espaciais, neste trabalho, denota conceitos ou representações mentais utilizadas pelos indivíduos para conhecer e classificar as instâncias semânticas dessas entidades ou eventos que ocorrem no mundo-real.

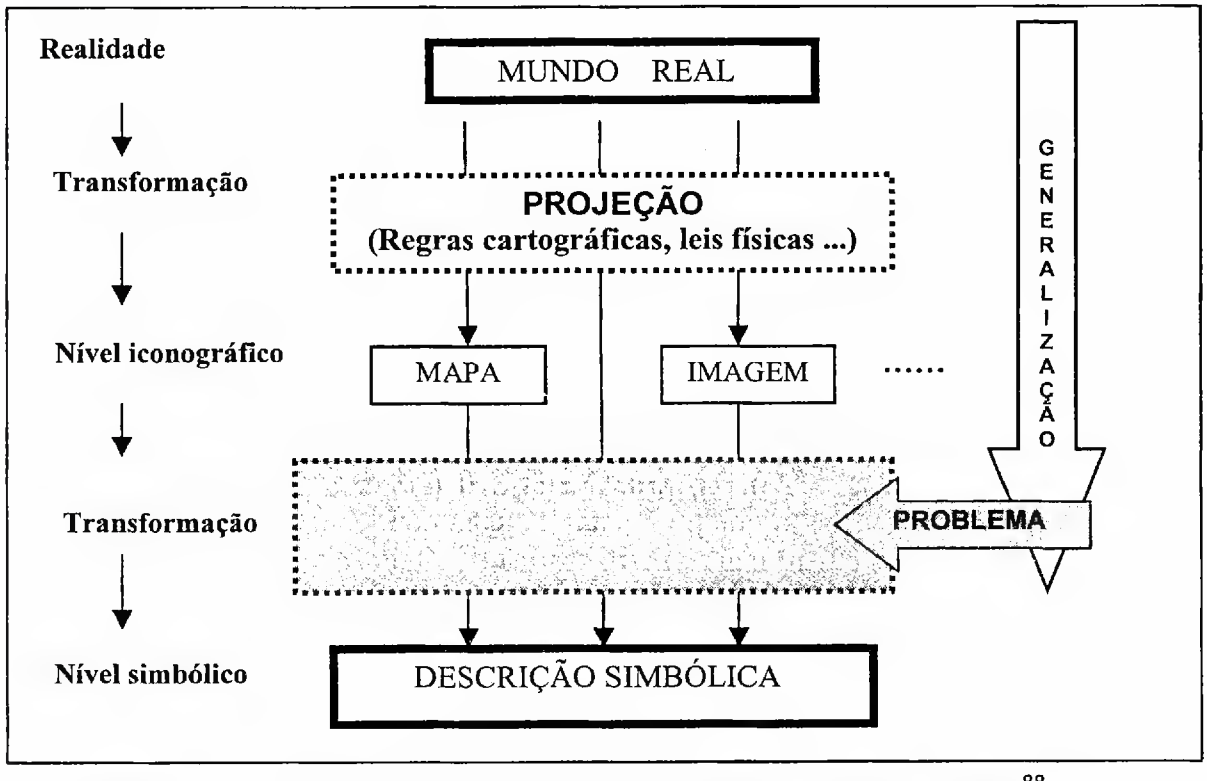

Figura 2.1: Níveis conceituais e transformações correspondentes.[88]

Esta tese é um prosseguimento da de RODRÍGUEZ (2000, p. 136-141), sem deixar de incorporar outras sugestões de trabalhos congêneres, como é o caso do de MEDEIROS (1999, p. 259) e do de BÄHR (1996, p. 9); todos os três, estudos que se enquadram na área de representação do conhecimento e do comportamento humano racional, i.e., estão no domínio da Inteligência Artificial (IA). Fora estes, que dão mais os limites teóricos da pesquisa, há outros trabalhos que balizaram o domínio mais prático da pesquisa, ligados à implementação do protótipo, os quais serão citados, oportunamente, nas seções de revisão de literatura e de metodologia.

[87] Relatadas em RODRÍGUEZ (2000, p. 2).
[88] Adaptada de BÄHR (1996).

2.2.2. Trabalhos precedentes

A Figura 2.1 despertou o interesse por esta área de pesquisa, que explora aspectos lingüísticos para representar a informação geográfica. BÄHR (1996), que a esboçou, estava preocupado em ***como se levantar fundamentos teóricos para representar as entidades espaciais do mundo-real num sistema de informações geográficas com a menor perda possível dos seus conteúdos semânticos***. Apesar de não ser esse o espaço-problema desta tese, pode-se considerar esta questão, tacitamente exposta por BÄHR (1996), como o embrião da questão geral de pesquisa que virá adiante.

A representação da informação espacial se torna crítica num ambiente computacional, de características binárias (verdadeiro ou falso, **0** ou **1**, certo ou errado). A natureza não é "binária". O problema de definição de limites nesse ambiente não é nada trivial, porque no mundo-real não existe definição exata de fronteiras ou limites (vale-montanha, rio-mar, etc.). Nessa linha de raciocínio, é claro que um processo que ficar além dessas injunções de máquina forneceria uma forte contribuição, tendo em mira integrar SIGs que hoje em dia ainda conservam enfoques limitados do mundo que pretendem modelar.

Nos níveis mais baixos de integração da informação, num sistema de informação, são as fracas e clássicas relações sintáticas que predominam; num nível acima, vêm os esquemas de bancos de dados; finalmente, no mais alto nível, encontram-se as definições semânticas, que oferecem o maior poder de integração e representatividade da informação (V. Figura 1.13). É aí que estão se formando as questões de pesquisa preocupadas com a construção de ontologias que sejam capazes de: 1) Captar significados relevantes numa base de dados, independentemente da forma como esses dados estejam organizados; 2) Propiciar buscas (*queries*) bem específicas (de conteúdo) do usuário a essa base, "descobrindo semelhanças onde aparentemente há diferenças e descobrindo diferenças onde aparentemente há semelhanças" (GORSKI, 1968).

Para continuar descrevendo os trabalhos que antecederam esta pesquisa, é preciso aduzir para este subitem algumas considerações sobre objetos[89] espaciais representados pelos seguintes itens:

- Relações espaciais que mantêm entre si: topológicas, de distância e as de direção;
- Características geométricas que lhes são peculiares: forma, tamanho e densidade;
- Atributos (propriedades semânticas): capazes de classificar os objetos espaciais.

Os três tipos de itens acima vêm na forma de volumosos acervos de dados que constituem as *cenas espaciais*[90], material de que os SIGs sempre trataram.

Os primeiros trabalhos relatados por RODRÍGUEZ (2000) nessa área, começaram em 1981, com as *análises espaciais* que os geógrafos realizavam sobre *conjuntos de pontos*.

Na metade da década de 90, começaram os estudos, propriamente ditos, sobre a similaridade de cenas espaciais pelo enfoque da *descrição visual* dos objetos espaciais (cor, forma, textura e tamanho). Havia uma outra tendência de estudos dessa linha, que optava pela avaliação da similaridade de cenas espaciais segundo o arranjo *ou disposição das entidades espaciais* entre si. Os critérios para o julgamento, nesse caso, são: direção (N, S, E, W), topologia (dentro, fora, à esquerda, etc.) e distância.

RODRÍGUEZ (2000) justificou a *natureza cognitiva* de seu estudo, alegando que a similaridade semântica (SS) não se refere às propriedades geométricas dos objetos que representam entidades do mundo-real, mas aos termos (nomes) que denotam essas entidades. Entendendo-se as relações conceituais que se estabelecem entre esses termos, pode-se entender as descrições que as pessoas fazem das entidades espaciais. Max Egenhofer [*apud* RODRÍGUEZ (2000)] denominou essa área de pesquisa de *Naive Geography* (Geografia do senso comum, do quotidiano ou ingênua), preocupada com a complexa matéria de formalizar conhecimento do dia-a-dia das pessoas sobre o espaço, de forma dinâmica.

Do mesmo modo que MEDEIROS (1999) fez em relação à recuperação de informação, RODRÍGUEZ (2000) citou diversos autores que começaram a trabalhar com a SS, enfrentando o grande problema da ambigüidade das línguas naturais.

O complexo problema da ambigüidade pode ser ilustrado com o exemplo da palavra "banco", nesta história hipotética: "O ladrão roubou o *banco* da cidade, fugiu de barco, mas o atolou no *banco* de areia do rio. Preso pela polícia, foi algemado ao pé do *banco* de cimento da praça, enquanto o policial consultava o *banco* de dados sobre a sua ficha criminal". É o contexto presente na língua natural, difícil de formalizar e implementar em máquina, que facilita o entendimento das várias acepções (polissemia) da palavra banco, nessa pequena história.

A ambigüidade de texto pode ser controlada pelos tradicionais tratamentos terminológicos no campo das ciências da classificação. A tendência nessa área, no entanto, parece pautar-se pela construção de sistemas de PLN (IA) dotados de bases de conhecimento que contenham conceitos genéricos ou específicos de certa área do saber, não se tendo certeza (no âmbito desta revisão de literatura) sobre o desempenho do método tradicional em relação ao de um baseado em IA. RODRÍGUEZ (2000), no entanto, assegurou que sistemas de

---

[89] Lembrar que o objeto é uma abstração que só existe quando uma entidade (coisa) do MR é representada numa BD.
[90] Fotos, imagens e mapas, notadamente.

PLN que tratem de áreas muito abrangentes do conhecimento "acabam por cair nas armadilhas da ambigüidade", além de ser inviáveis operacionalmente em termos de esforço computacional (problema NP[91]). Daí a autora ter frisado que o seu trabalho de pesquisa foi limitado ao campo de conhecimento atinente ao domínio geográfico. Por extensão, esta também foi a linha de orientação seguida nesta tese.

Os aspectos limitativos enumerados por RODRÍGUEZ (2000, p.13) para esses trabalhos, também presentes no primeiro modelo se SS desta pesquisa, implementado por um protótipo chamado de PROFAX, são os seguintes:

- Dependência contextual;
- Avaliação de assimetria;
- Representação adequada do sistema de conceitos espaciais.

2.2.3. A natureza do problema de pesquisa

As referências de julgamento da similaridade entre objetos espaciais que "povoam" uma base de dados foi o objetivo central da tese de doutorado da qual esta pesquisa se originou: RODRÍGUEZ (2000). Esta autora se utilizou de critérios subjetivos e mesmo intuitivos, ligados a processos cognitivos da mente humana. O protótipo[92] de um sistema de mensuração dessa similaridade foi implementado numa linguagem de programação de alto nível (C++), para provar que é possível criar um instrumento que simule critérios subjetivos de julgamento de similaridade do ser humano. Tais processos prometem revolucionar a modelagem dos SIGs, já se falando em SIGAIAs (Sistemas de Informação Geográfica Apoiados pela IA[93]).

Repetindo, a meta final desses esforços de pesquisa está na *interoperabilidade* entre SIGs, passando pelo desenvolvimento de interfaces amigáveis (*interação*) com um usuário não especializado em *geoprocessamento*. Como disse EGENHOFER (2001b): "Não se trata apenas de um trabalho de engenharia, mapeando uma especificação de *software* para outra. Por trás, há aspectos complexos de *integração* desses *softwares* que transcendem a sintaxe"

Com isso, EGENHOFER (2001b) quis dizer que é preciso começar a integração devagar, em pequenos passos, escolhendo-se um domínio específico de problema e de *softwares*-solução. O problema de integração, no entanto, torna-se quase intratável quan-

---

[91] V. glossário.
[92] Chamado por RODRÍGUEZ (2000) de *MD (Matching-Distance) model.* Aqui será chamado de MSS: modelo de similaridade semântica.
[93] KBGIS: *Knowledge Based Geographic Information Systems* (Univ. da Califórnia – Santa Bárbara).

do são adicionadas novas funções e capacidades aos *softwares*-solução. A causa das dificuldades, pelo menos, já foi detectada: está no domínio semântico das diversas implementações. O componente semântico da IG é a chave da interoperabilidade dos SIGs do futuro, mas é um item crítico, que só pode ser encarado com sólida fundamentação teórica de Engenharia de *Software*. O assunto tem vindo à tona desde meados da década de 90 (DAWN, 2002) e nos encontros COSIT (*Conference Series on Spatial Information Theory*).

No caso da Engenharia de *Software*, PRESSMAN (1995) relatou que algumas técnicas de apoio ao desenvolvimento de *software* nessa área (CASE) faziam pouco uso das tecnologias de IA. Na maior parte das vezes, a IA era empregada para verificar a exatidão gráfica dos modelos de análise e de projeto. No entanto, já naquela época, como continuou o autor, a perspectiva das pesquisas na área eram as de usar as ferramentas CASE-AI em agentes "inteligentes" para auxiliar na análise e projeto de sistemas além da simples conferência de gráficos, incluindo a fase de testes de sistemas.

Sendo assim, PRESSMAN (1995) prognosticara que tais AIs prometiam um futuro produtivo, restando apenas, àquela época, a formação de bases adequadas de conhecimento de Engenharia de *Software*.

Hoje em dia, a literatura digital na rede mundial confirma o prognóstico de PRESSMAN (1995), porquanto registra o aparecimento de uma ferramenta chamada SIS (Sistema de Informação de *Software*), desenvolvida para isolar os problemas de representação inerentes às linguagens de programação e auxiliar na montagem de uma base de conhecimentos, empregando mecanismos de busca de informações relativas ao *software*. Tais informações podem ser explícitas ou não; neste caso as informações são sensíveis ao contexto de uso desse *software*, tarefa nada trivial de modelar.

Essa literatura esparsa sobre IA e CASE denomina esses sistemas especialistas SIS de CSIS (Sistema de Informação de *Software* Compreensivo), quando as metas traçadas para o sistema original atingem um grau razoável[94] de produtividade.

Portanto, resta um esforço conjunto e concentrado, em foros adequados, para definitivamente iniciar a construção de SIGs semânticos. Como EGENHOHER (2001b) já explicara, trata-se de fundamentar teoricamente e carrear um razoável aporte de observações para a área de estudos cognitivos da Geografia do Quotidiano (*Naive Geography*), acessível a um bem amplo espectro de usuários e inaugurar a etapa mais almejada pelo OGC, no que tange ao intercâmbio de dados e informações cartográficas digitais.

---

[94] O emprego do novo conceito de ontologia nesses sistemas é fundamental para este resultado.

EGENHOFER (2001b) argumentou que essas metas só serão alcançadas com uma mudança de enfoque na fase de modelagem desses novos SIGs semânticos, consistindo em se entender o que há por trás das diferentes estruturas geométricas com que foram concebidos os SIGs tradicionais. As *geometrias*, como ele disse, "são aspectos importantes mas não essenciais na modelagem de SIGs semânticos; o que importa é entender as diferentes abstrações com que os seus criadores 'enxergaram' o mundo-real. Uma comunicação bem-sucedida tem suas bases no mesmo modelo mental que emissor e receptor devem possuir sobre o conteúdo da mensagem que flui entre ambos.

Nos domínios da interoperabilidade, a comunicação bem-sucedida significa dizer que há dois ou mais emissores (coletores de dados) e um receptor (usuário do SIG) e que este receptor deve ser provido por este sistema de recursos para *integrar* as informações advindas de diversas fontes de dados.

Já foram mencionados *en passant* os termos ***recuperação*** e ***integração de informação***, que podem ter deixado alguma brecha de entendimento, apesar de que, no contexto em que foram empregados, uma vaga noção dos requisitos para ambos pode ter ficado: ***precisão*** para o primeiro e ***compatibilidade*** para o segundo.

No que tange à *recuperação da informação*, os subitens 1.2.3 e 1.5.1 já deram uma idéia do estado-da-arte sobre o assunto. Neste subitem, contudo, é interessante colocar o assunto na perspectiva da similaridade semântica, como fez RODRÍGUEZ (2000).

Nos sistemas de informação tradicionais, os usuários exprimiam as suas necessidades por informação por meio de consultas, que poderiam basear-se num conjunto de combinações *booleanas* de palavras-chaves, em sentenças de língua natural ou em interfaces gráficas (*user-system dialogs*). Com o avanço das TIs, os sistemas passaram a tratar de tipos mais diversos de informação em suporte digital, além de texto (som, imagens, mapas, etc.), provocando um crescente interesse por novas formas de interface com os usuários. Uma capacidade desejada para estas foi a de proporcionar ao usuário a possibilidade de recuperar a informação do seu interesse sem ter que se preocupar com as estruturas internas de dados do sistema. Essa é uma características das linguagens de manipulação de dados, encontradas em bancos de dados, em que o acesso lógico aos dados é separado do acesso físico.

Em virtude dos imensos repositórios de dados que foram se formando desde a década de 70 e pelo fato de que as estruturas de armazenamento desses dados poderia não refletir a natureza real da informação neles contida, não seria plausível esperar que os usuários

tirassem vantagem completa na utilização desses sistemas em suas consultas. Daí a necessidade de aprimoramento nesta área de recuperação da informação.

RODRÍGUEZ (2000) e MEDEIROS (1999) definiram a ***recuperação de informação*** como uma operação que combina (casa ou coincide) da melhor forma possível o conteúdo (termos) da consulta do usuário com o conteúdo informativo representado internamente numa base de dados. MEDEIROS (1999), a propósito, ressalta o fato de que a recuperação de informação relevante para o usuário final, que se utiliza de um sistema de informação, é uma área prolífera de pesquisa em Ciência da Informação.

Num passado recente, o processo de "casamento" (*matching*) da maioria dos enfoques dos SRIs (sistemas de recuperação de informação) fundamentava-se em análise estatística de termos indexados. Em breve esses sistemas chegaram ao seu limite, uma vez que eles eram mais dependentes de correspondências sintáticas (forma) do que de correspondências semânticas (significado, conteúdo) na recuperação da informação (RODRÍGUEZ, 2000). O exemplo a seguir ilustra como a descoberta de informação relevante é tratada por um SRI de fundamento semântico, retratando a natureza do problema dessa linha de pesquisa de uma forma mais realista.

Se um usuário for consultar uma base de dados geográficos à procura cidades no estado de São Paulo que possuam pelo menos uma universidade, é provável que num sistema de natureza sintática, as restrições extremas de forma deixem de lado centros universitários ou faculdades isoladas, que poderiam estar no domínio de interesse desse usuário. Uma busca semântica transcende essas restrições de forma, não se detendo em fazer exatas coincidências (*matchings*) entre o que o usuário formalizou no seu pedido de busca e o conteúdo da base de dados. Com base numa conjunto de regras que implementam a similaridade semântica entre termos como *universidade*, *centros universitários* e *faculdades isoladas*, um sistema mais "inteligente" poderia oferecer um rol maior de opções de busca para o usuário e aumentaria as suas expectativas de satisfação. Dessa forma, a similaridade semântica é uma espécie de ferramenta que contribui efetivamente na busca exploratória por informações (que implica descoberta), já que o usuário não sabe exatamente, de antemão, pelo que está procurando.

Nesse ponto, é bom distinguir *localizar dados* de *descobrir informação*. Entende-se por descoberta da informação a localização de objetos de interesse dentro de uma população de objetos potencialmente relevantes e distribuídos de várias maneiras, onde a natureza da distribuição pode variar entre o caótico e o altamente organizado. Na simples

localização, não existe esse fator de relevância[95] dos objetos. A busca é feita de maneira determinística (ou coincide ou não coincide). SRIs avançados contemplam algoritmos associados à descoberta da informação (não mais localização). Localizados os documentos (textos, mapas), a recuperação dos seus conteúdos é mais abrangente (descoberta), se estiver presente um processo de análise semântica - *parsing* (BORGES, 2002).

No que tange à ***integração da informação***, sintetizando RODRÍGUEZ (2000), há uma diferença primordial em relação à integração de dados, porque a primeira está num nível superior de processamento, selecionando as informações necessárias ao SIG de uma base de dados. As razões da preocupação com a integração de dados estão no surgimento de ambientes distribuídos (Internet), que exigem constante reutilização[96] e compartilhamento desses dados.

A heterogeneidade desses acervos é o maior desafio nessa área. As soluções passam pelo entendimento de três arquiteturas de ambientes de dados distribuídos: 1ª) Esquemas globais; 2ª) Sistemas de bases de dados federativas (FDBS) e 3ª) Linguagens para agrupamentos de bases de dados. O que importa é que a variável independente da função de integração desses esquemas está no rigor da coesão entre as bases de dados de cada um: mais rigor implica mais estabilidade na integração das bases, com perda no aspecto da flexibilidade na sua administração (usuário longe do controle, administrador no controle).

Segundo FRANK (1999), GAHEGAN (1999) e SHETH (1999), como a recuperação, a integração de informação está intimamente ligada ao problema de SS (e à interoperabilidade).

Sintaxe – esquema da base de dados (identificação de entidades no MR) – *semântica* (*relações entre as entidades*): esta é a ordem de abstração, do mais formal ao mais conceitual, para se alcançar a meta de interoperabilidade. Tal ordem caracteriza a evolução dos SIGs. A maioria dos SIGs atuais foi desenvolvida pelo enfoque intermediário de abstração. O desafio é o semântico (SHETH, 1999).

A Figura 2.2 mostra o esquema geral da integração de informações em SIGs de concepção semântica.

O problema da SS está na identificação de objetos semanticamente similares e que pertençam a diferentes arquiteturas de base de dados, assim como na resolução de suas diferenças esquemáticas. Nem todos os dados das bases já formadas podem ser integra-

[95] V. glossário.
[96] A linguagem Java™ é muito promissora neste requisito.

dos e, por enquanto, a SS é uma das maneiras capaz de promover a integração dos casos possíveis, conforme registrou RODRÍGUEZ (2000), que ainda analisou diversos estudos e propostas *ad-hoc*, assinalando a que é baseada em definições e inter-relacionamento entre os termos descritores de entidades do MR numa base de dados (BD).

O cerne desta proposta está em produzir um mapeamento entre os termos da BD e uma ontologia construída, tal que capturasse a visão do mundo segundo a utilização que se quer dar àquela BD. Tal ontologia deveria ter as seguintes características:

- Propiciar consultas *intensionais*[97] (conseqüência da finalidade primordial da ontologia: atender a um grupo específico de usuários);
- Definir o nível semântico do sistema independentemente do nível de representação de dados (esta é uma métrica tradicional da Engenharia de *Software*);
- Reproduzir a relevância (V. glossário) dos dados, sem precisar ter acesso a eles.

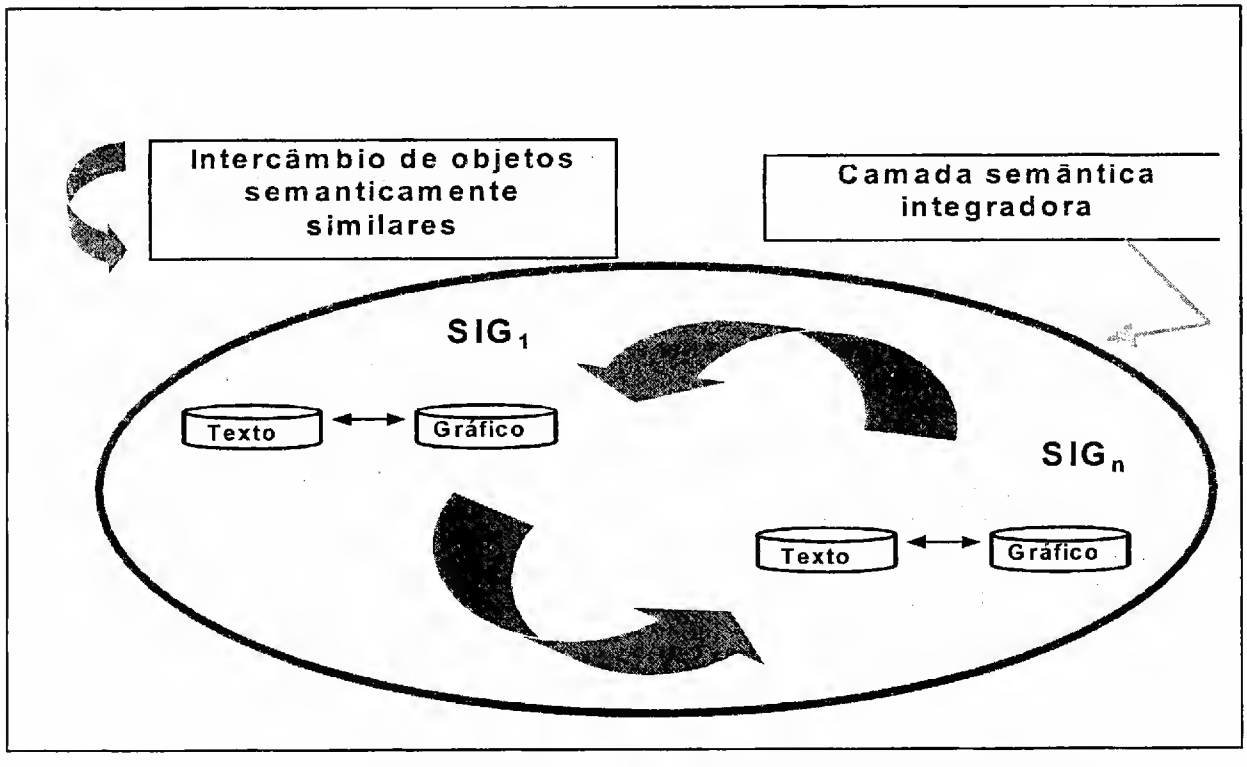

Figura 2.2: A integração de informações nos futuros SIGs.

Definida uma ontologia comum, os relacionamentos entre os termos expressos pelas sentenças lógicas da ontologia são transformados em similaridades semânticas.

---

[97] Relativas ao conteúdo informativo da BD. Alguns autores da Ciência da Computação desconhecem o termo com "s".

Ao se criar uma ontologia comum para SIGs, inicia-se a construção de uma base de conhecimentos (BC), que vai estar pronta para ser ampliada com mais conhecimento oriundo das verificações das informações obtidas das bases de dados com um conjunto de axiomas básicos da BC. RODRÍGUEZ (2000) relata projetos-pilotos nesta área, na Europa e nos EUA. É, portanto, muito importante começar esforços similares no Brasil.

Ao se falar de integração de SIs, não se pode deixar de pensar na maneira que vem se mostrando mais promissora para explicitar (explanando) um plano de integração desses SIs: por intermédio de uma *ontologia*! E, no atual contexto deste trabalho, o que vem a ser uma **ontologia**?

A definição operacional, a seguir, é uma síntese de RICH (1993), RUSSELL (1995), MEDEIROS (1999) e FONSECA (2000b) e procura delimitar o problema geral, que logo será formulado.

Ontologia é um conjunto de poucas *regras* gerais (axiomas), formalizadas por enunciados claros e concisos, que descrevem uma certa realidade em termos de um conjunto de classes de objetos e de suas *inter-relações* (hierarquia). Em SIs modernos, as ontologias "capturam" a semântica das fontes de dados e são a base da recuperação e da integração de informação. A estrutura de uma ontologia deve ser capaz de inferir questões (*queries*) de busca de informação (*recuperação*) semanticamente ricas, a ponto de propiciar a melhor conformação possível entre essas questões de busca e os dados contidos em BDs (distribuídas ou não). Além disso, a estrutura de uma ontologia deve manter uma base de conhecimentos compartilhada, num ambiente de SIs heterogêneos, permitindo o intercâmbio de objetos semanticamente similares (*integração*). As ontologias serão os "cérebros" dos *web crawlers* (robôs de busca da *web*) dos próximos anos.

A definição anterior para ontologia, se interpretada de forma abrangente, engloba até mesmo um algoritmo. Segundo a vasta literatura de Ciência da Computação, um algoritmo é a descrição de um padrão de comportamento, expresso em termos de um *repertório* bem definido e finito de *ações*. No caso específico desta pesquisa, os algoritmos foram dotados de ações para criar e percorrer uma *taxinomia* (hierarquia), que traduz parte do modelo conceitual de uma BD orientada a objeto, bem como de ações para executar um *modelo matemático* que avalia a similaridade semântica entre as classes de entidades espaciais representadas pelos termos nodais da taxinomia.

Portanto, intrinsecamente, os códigos em *Java*™ dos dois protótipos desenvolvidos nesta pesquisa podem ser considerados como *ontologicamente* orientados. É possível perceber isto apenas pela comparação dos termos grifados do parágrafo anterior com a

definição de ontologia do penúltimo parágrafo. Os comentários a seguir, além de assinalarem tais analogias, acrescentam mais alguns traços peculiares da especificação dos protótipos, o que já sinaliza para a sua categorização como LRCs (linguagens de representação do conhecimento):

- "Regras gerais da ontologia" e "repertório de ações" do protótipo. Ressalva: as regras algorítmicas[98] são muito mais simples que as sentenças lógicas que irão constituir os axiomas da BC de um protótipo ontológico. No entanto, o PRONTO©, além do seu código-fonte, em si já coerente com a definição de ontologia, ainda contém axiomas de natureza totalmente ontológica, destinados a explicitar conceitos e que foram transformados em sentenças declarativas de uma LTP OO como o *Java*™;
- "Inter-relações hierárquicas" da ontologia e a "taxinomia" dos protótipos;
- "Recuperação" da ontologia e, de novo, "repertório de ações" do protótipo. Ressalva: uma ontologia implementada é um passo além em matéria de capacidades (semânticas) de busca "inteligente" de informação, em relação aos tradicionais métodos algorítmicos (método de seleção, da bolha, da ordenação, etc.).

Aproveitando esta última ressalva, vale registrar que até mesmo o modelo de SS implementado pelo PROFAX não está destituído de potencial semântico. Ele utiliza o co-seno entre dois vetores do espaço euclidiano para reproduzir a SS entre classes de entidades espaciais. A semântica do processo está por trás das propriedades do co-seno entre dois vetores, os quais são agentes de representação de termos, que, por sua vez, representam classes de entidades espaciais do MR. Essas transformações sucessivas de coisas (fatos) de mundos distintos, ou por meio de representação mental (entidade espacial em termo), ou pela mediação lógico-matemática (termo em vetor), constituem-se em verdadeiras "ondas" portadoras de significado, de informação útil (ou não[99]), capaz de produzir conhecimento, evidentemente de essência mais simples que o produzido pelo MSS de RODRÍGUEZ (2000) ou pelo modelo implementado pelo PRONTO©, que são de base genuinamente ontológica.

Dentro desse contexto, o que mais diferencia o presente trabalho do de RODRÍGUEZ (2000) é o objetivo geral da autora, que é muito mais abrangente. O MSS de RODRÍGUEZ (2000) possui uma visão de integração semântica entre diferentes SIGs, por meio da avaliação de SS entre termos de diferentes ontologias (*cross-ontology evaluation*), enquanto que o PRONTO© restringiu-se a um estudo-de-caso sobre um subconjunto

[98] Trinômio seqüência – alternativa(ou decisão) – repetição de BÖHM (1966).
[99] Se a transformação foi inadequada e degradou o conteúdo original da informação, prejudicando o entendimento.

do MC da folha Faxinal, ou seja, os termos designativos das classes de entidades espaciais deste modelo foram explicitados numa ontologia isolada.

2.2.4. A formulação do problema de pesquisa

O enunciado do problema geral de pesquisa originou-se da leitura do artigo de BÄHR (1996), primeiro documento do levantamento bibliográfico.

Por intermédio da investigação das referências bibliográficos deste autor e pela via cronológica, de 1996 até 2000, chegou-se ao trabalho de RODRÍGUEZ (2000), ao se apelar, de modo intensivo, para o serviço www da Internet no que tange a buscas em serviços especializados, pedidos de auxílio a listas de assunto e à subscrição em grupos de comunicação científica.

Os tópicos a seguir, retirados na íntegra do artigo de BÄHR (1996) e da tese de RODRÍGUEZ (2000), dão o tom da composição da formulação do problema geral:

Reclamo de BÁHR (1996)[100]:

"... transformations from the real world, through iconic level to symbolic level are subject to imprecisions. No rigorous theory of this type of error propagation has yet been formulated, but such a theory would be a prerequisite for semantic modelling..."

Propostas de trabalhos futuros (RODRÍGUEZ, 2000):

"Future work may consider an implementation of the semantic similarity model that uses a formalism for expressing structured and sharable knowledge ... Description logic gives a logical basis for frames, semantic networks, and object-oriented representations as well as for semantic models."

Outros trechos dessa autora, além de prover informação útil para o desenvolvimento de um dos instrumentos de validação da hipótese de pesquisa, o PRONTO©, pôde ser reaproveitado na recomendação para trabalhos futuros, estendendo o alcance das conclusões da autora e as que forem tiradas no Capítulo 7 deste trabalho. São estes os trechos:

"A further study, however, should examine whether or not the performance of the MD model under the same set of evaluations is better than the performance of existing models. Such as study could lead to the conclusion that the different approaches provide complementary answers and that no single model, but multiple approaches to semantic similarity should be considered depending on the semantic organization of entity classes.

---

[100] V. Figura 2.1.

Ontology vs. Database Schema: ...Ontologies and database schemas are related, but not equivalent. Ontologies have explicit representations of the entity classes' semantics, whereas database schemas usually use implicit semantics ...Entity classes in a database could be associated with their corresponding ontological definitions through a semantic directory[101]*. The creation and maintenance of these directories are areas for further research as well as solving schematic conflicts that are product of different levels of abstraction in the entity class representations ."*

Problema geral de pesquisa de RODRÍGUEZ (2000):

"What are the desirable properties of a similarity model among spatial entity classes?
What are the main components that semantically distinguish spatial-entity classes?
What are the advantages and disadvantages of current models for semantic similarity?
Can advantages of current models be integrated into a new similarity model?
How does context affect similarity assessment?"

Concluindo parcialmente sobre os excertos do artigo de BÄHR (1996) e da tese de RODRÍGUEZ (2000), é possível explicitar um enunciado de problema geral:

"Como avaliar a similaridade semântica entre objetos espaciais representáveis numa base de dados?"

Além da influência dos dois autores citados, as concitações para trabalhos futuros de MEDEIROS (1999) também tiveram peso na formulação do problema geral. A seguir, os excertos relevantes das concitações de MEDEIROS (1999, p. 259):

"As ferramentas desenvolvidas nesta tese podem ser aperfeiçoadas e podem ser aplicadas a outras pesquisas ... como Tradução Automática, Lingüística, Terminologia. Neste sentido, sugerem-se: ...

c) Aplicação da metodologia de ... tratamento semântico-sintático em corpus de outras áreas do conhecimento."

Há outros trabalhos mais assemelhados a esta pesquisa do que os relacionados no subitem 2.2.2, sobre os quais MEDEIROS (1999) e RODRÍGUEZ (2000) fazem referência. Esses trabalhos serão examinados na revisão de literatura, onde todos serão integrados de forma a proporcionar o embasamento para a enunciação da hipótese de pesquisa e para o estabelecimento das formulações matemáticas de cálculo da SS entre objetos espaciais.

[101] Em "Pontos e Contrapontos", há uma troca de correspondência com a autora, esclarecendo o que significa "semantic directory" entre outras coisas.

2.3. Objetivo geral da pesquisa

Concentrando tudo o que foi até agora descrito sobre o interesse original por essa área de investigação científica, os trabalhos precursores, a natureza do problema geral e, finalmente, a própria forma de apresentação do problema geral, segue-se o enunciado do objetivo geral que norteará a metodologia de pesquisa:

"Avaliar a similaridade semântica entre objetos espaciais representáveis numa base de dados."

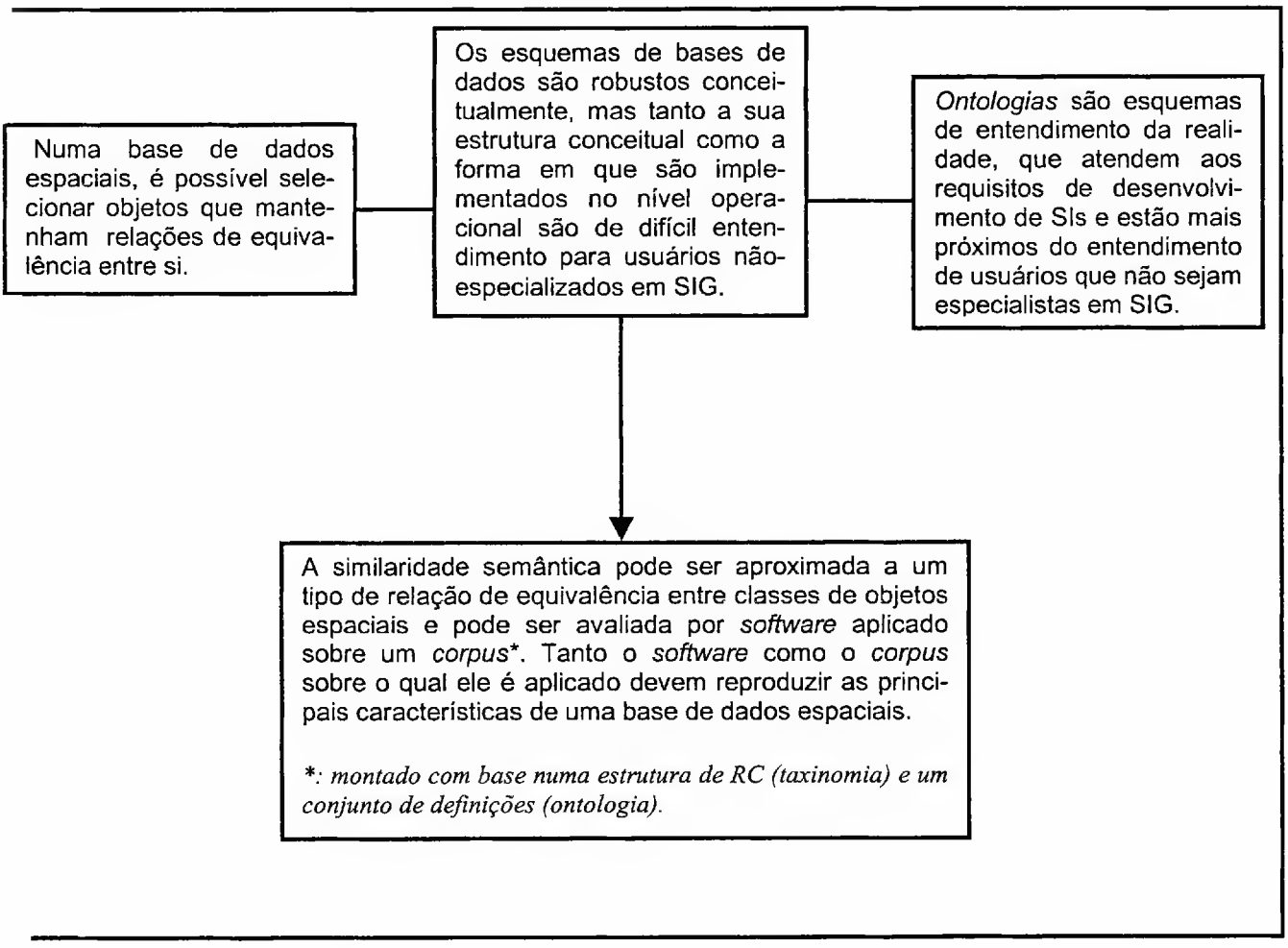

Figura 2.3: Base lógica da pesquisa.

Para concatenar logicamente as principais idéias que decorrem das questões e dos objetivos gerais desta investigação exploratória, foi montado um quadro (Figura 2.3), que ainda teve a finalidade de ser um guia na revisão de literatura, de juntar as peças realmente necessárias de orlar uma satisfatória hipótese de pesquisa. Para o leitor, fica o cenário pano-

râmico dessas seções introdutórias da tese, brevemente descritas no próximo subitem, que encerra o capítulo.

2.3.1. Base lógica (análise racional[102])

O esquema da Figura 2.3 ilustra a base do raciocínio desta proposta de pesquisa, já se delineando no retângulo inferior a hipótese geral do trabalho.

Para a consecução dos objetivos já expressos e para manter firme a base lógica esboçada na Figura 2.3, a revisão de literatura deverá estreitar a sua busca nos seguintes campos do conhecimento: 1) Estruturas de representação do conhecimento da IA; 2) Teoria do Conceito e técnicas para definir classes de objetos espaciais similares, para montar um *corpus* que privilegie relações todo-parte e gênero-espécie entre essas classes; e 3) Teoria e prática na área de SIG, para gerar as unidades de observação e variáveis.

[102] Artifício aproveitado de LOCKE (1993, p.10-12); no original: *rationale*.

3. REVISÃO DE LITERATURA

"O espaço é um conjunto de objetos e de ações ... Geometrias não são geografias ... É preciso reinterpretar a lição dos objetos que nos cercam e das ações das quais não podemos escapar." (Milton Santos)

3.1. Revisão de literatura da área de Cartografia

Como o objetivo geral desta pesquisa possui um caráter interdisciplinar, a revisão de literatura, em conseqüência, incorpora o espírito cooperativo entre as duas grandes áreas categorizadas para esta revisão: Cartografia e ciências cognitivas.

A presente tese não constitui um esforço exclusivo de investigação sobre assuntos técnicos que a Cartografia suscita, notadamente no domínio das engenharias. A intenção desta revisão em Cartografia foi justamente a de deslocar o aparente papel de protagonista das *geociências*, de forte caráter quantitativo e experimental, para o domínio interdisciplinar que os futuros SIGs já começam a palmilhar, importando conceitos da Psicologia, da Neurofisiologia, da Filosofia, da Lingüística, da IA, enfim, das ciências cognitivas.

Dessa forma, o leitor poderá identificar os diversos elos de ligação deste tópico da revisão com o próximo[103], o que o torna uma espécie de preâmbulo (prolegômenos) do seguinte.

O item que for de natureza muito técnica não será aprofundado, uma vez que não constitui o objetivo a ser alcançado pela metodologia e, dentro do possível, já recebeu a adequada descrição no glossário.

3.1.1. Estado-da-arte do *Geoprocessamento*

Apreciar o quadro sinóptico das tecnologias de *Geoprocessamento* no tempo, bem como visualizar os cruzamentos que manteve com outras tecnologias, é fundamental para compreender a extensão desta pesquisa e os sistemas conceituais que começaram a se consolidar a partir da década de 90 e de cuja origem o *geoprocessamento* foi a matriz.

Não se deve perder de vista um freqüente fato histórico no desenvolvimento científico: a contribuição dos saltos no campo tecnológico. Um instrumento de medição mais preciso, um engenho inventado para proporcionar mais comodidade à sociedade; todos esses fatos, de uma forma ou de outra, trouxeram dividendos para as teorias científicas, mesmo que tenham começado de forma empírica e até improvisada. Pode-se dizer que a técnica é cega

[103] Ciências cognitivas.

sem a ciência. Por outro lado, a ciência é estéril sem a técnica. O ideal é que ambas andem juntas (HUISMAN, 1976).

Nos dias de hoje, não se pode falar em Cartografia se não se falar em *geoprocessamento*, termo denotativo de um conjunto de várias técnicas, muitas das quais já foram transformadas em disciplinas acadêmicas, como SIG, p.ex.

No *cenário internacional*, as primeiras tentativas de automatizar parte do processamento de dados com características espaciais aconteceram na Inglaterra e nos Estados Unidos, na década de 50, com o objetivo principal de reduzir os custos de produção e manutenção de mapas. Dada a precariedade das TIs da época e a especificidade das aplicações desenvolvidas (pesquisa em Botânica, na Inglaterra, e estudos de volume de tráfego, nos EUA), esses sistemas não podem ser classificados como "sistemas de informação" (CÂMARA, 2002).

Como já citado, os primeiros SIGs surgiram na década de 60, no Canadá, como parte de um programa governamental para criar um inventário de recursos naturais. Tais sistemas, no entanto, eram muito difíceis de usar: não existiam monitores gráficos de alta resolução, os computadores necessários eram excessivamente caros e a mão-de-obra tinha que ser altamente especializada. Não existiam soluções comerciais prontas para uso e cada interessado precisava desenvolver seus próprios programas, o que exigia muito tempo e, naturalmente, muito dinheiro. Além disso, a capacidade de armazenamento e a velocidade de processamento eram muito baixas.

Na década de 70, foram desenvolvidos novos e mais acessíveis recursos de *hardware*, tornando viável o desenvolvimento de sistemas comerciais. Foi nesta época que a Engenharia de *Software* começou a se firmar, não deixando de beneficiar o campo da Engenharia tradicional com as técnicas de projeto apoiado por computador (CAD), que melhoraram em muito as condições para a produção de desenhos e plantas para a Engenharia Civil e Mecânica, além de servirem de base para os primeiros sistemas de cartografia automatizada. Também na década de 70, foram desenvolvidos alguns fundamentos matemáticos voltados para a Cartografia, incluindo questões de Geometria computacional. No entanto, em razão dos custos e pelo fato de estes primeiros sistemas ainda utilizarem exclusivamente computadores de grande porte, apenas grandes organizações tinham acesso à tecnologia.

A década de 80 do séc. XX representou o momento de acelerado crescimento nas tecnologias ligadas a SIGs, que dura até os dias de hoje. Até então limitados pelo alto custo do *hardware* e pela pouca quantidade de pesquisa específica sobre o tema, os SIGs se benefi-

ciaram grandemente da massificação causada pelos avanços da microeletrônica, da Informática[104] e do estabelecimento de centros de estudos sobre o assunto.

Nos EUA, a criação dos centros de pesquisa que formam o NCGIA[105] marca o estabelecimento do *Geoprocessamento* como disciplina científica independente. É nesse período que sobrevêm a grande popularização e o barateamento das estações de trabalho gráficas (*workstations*). Dois eventos dessa época concorreram para uma inigualável popularização dos SIGs: a invenção do computador pessoal (IBM-PC, em 1981) e o aprimoramento da interface gráfica dos sistemas gerenciadores de bancos de dados relacionais (SGBDs). Assim, usuários quase leigos em Cartografia e em Informática, estimulados pelas quedas nos custos de *hardware* e *software*, passaram a utilizar os SIGs para resolver problemas inerentes às suas necessidades e de maneira cada vez mais fácil. Essa tendência continua nos dias de hoje e de forma mais intensa (CÂMARA, 2002).

No *cenário nacional*, foi a iniciativa de pesquisadores do Departamento de Geografia da UFRJ, no início da década de 80, que permitiu a introdução do *Geoprocessamento* no ambiente acadêmico brasileiro. O marco lançado por aquela equipe de pesquisadores foi o desenvolvimento do sistema SAGA (Sistema de Análise *Geoambiental*).

Outros sistemas similares ao SAGA surgiram ainda na década de 80 e na década de 90, tanto da iniciativa privada como da governamental. Essa proliferação de SIGs e os problemas associados ao requisito de interoperabilidade já foram alvo de considerações anteriores. Por conseguinte, para não se afastar do ambiente acadêmico, mais comprometido com este requisito, acrescentam-se a este acervo as seguintes iniciativas brasileiras:

- **INPE**: em 1984, foi criado um grupo específico para o desenvolvimento de tecnologia de *geoprocessamento* e Detecção Remota (DR): a Divisão de Processamento de Imagens (DPI). De 1984 a 1990, a DPI desenvolveu o SITIM (Sistema de Tratamento de Imagens) e o SGI (Sistema Geográfico de Informações), para ambiente PC/DOS. Até 1991, o SITIM/SGI foi o suporte de um conjunto significativo de projetos ambientais no Brasil. A partir daí, este SIG foi substituído pelo SPRING (Sistema para Processamento de Informações Geográficas), para ambientes UNIX e MS/Windows, cujo sucesso, tanto em organizações públicas como privadas, advém da sua capacidade de unificar o tratamento de imagens de DR (ópticas e microondas), mapas temáticos, mapas cadastrais, redes e

[104] Outros desdobramentos desse avanço vertiginoso serão examinados no subitem 3.2.1.
[105] *National Centre for Geographical Information and Analysis*

modelos digitais do terreno (MDT). A partir de 1997, até hoje, mais uma vantagem do SPRING: ele passou a ser distribuído gratuitamente[106] pela Internet.

- **IME**: a partir de 1988, o Departamento de Cartografia desse instituto lançou a linha de pesquisa de SIG, com ênfase nas funções de saída (exibição), denominando-a de Sistema de Informações Cartográficas (SIC/IME). As dezenas de dissertações de mestrado e artigos publicados sobre este sistema contribuíram em aplicações tanto na área militar como na civil. Na área militar, o SIC/IME foi o embrião de aprimoramento de sistemas de guerra eletrônica, sistemas de armas e nas simulações de jogos-de-guerra. Na área civil, a cultura de SIG que foi disseminada pelo SIC/IME foi o marco teórico do projeto mais abrangente da DSG: representar o espaço geográfico do Brasil pelas tecnologias de modelagem de objetos (TMO), que disponibilizará para toda a sociedade produtos do mapeamento sistemático com mais rapidez, confiabilidade e menores custos. O projeto inicial, denominado de TBCD© (Tabelas da Base Cartográfica Digital), que contém oito categorias de entidades espaciais e centenas de feições gráficas, já está terminado. O manual técnico T 34-700 (BRASIL, 1998a), em parte derivado do projeto TBCD©, já foi aprovado e publicado. Esse manual é uma das fontes de dados para a formulação das definições das classes das entidades espaciais que fazem parte do *corpus* desta pesquisa.

3.1.2. Fundamentos epistemológicos da Ciência da Informação Geográfica

Um dos resultados da revisão de literatura neste tópico foi o de testificar o surgimento de um novo campo do conhecimento *geocientífico*, que vem sendo denominado por alguns meios de Ciência da *Geoinformação* e por outros de Ciência da Informação Geográfica (**CIGeo**). Nesta obra, o segundo termo será o acolhido, para manter a coerência com o principal trabalho de pesquisa do qual este se deriva: a tese de doutorado de RODRÍGUEZ (2000), da Universidade do Maine, que possui um dos centros de pesquisa do NCGIA.

O livro de CÂMARA (2002) é a prova de que o mesmo processo se dá no Brasil.

3.1.2.1. Da necessidade de conceitos

Ao contrário de outras disciplinas como Banco de Dados, p.ex., as TIs de *geoprocessamento*, particularmente de SIG, não possuem um corpo básico de conceitos, mas uma diversidade (por vezes contraditória) de noções empíricas.

---

[106] Vale assinalar que sistemas desse gênero, no mercado, não ficam por menos de dois mil reais.

De MOLENAAR (1991) até DAWN (1997), vários foram os trabalhos que clamavam por uma teoria sobre a informação geográfica. Todos esses trabalhos precursores contribuíram de alguma forma, introduzindo definições, identificando problemas e tecendo considerações de ordem epistemológica sobre o assunto. No Brasil, a trilogia organizada por CÂMARA (2002) pode ser considerada como o divisor d'água entre empirismo e teoria sobre a informação geográfica. Essa trilogia é composta das seguintes partes: 1ª) Introdução à Ciência da *Geoinformação*[107]; 2ª) Análise Espacial; 3ª) Bancos de Dados Geográficos.

Muitos livros-textos e cursos são organizados e apresentados em função de um sistema específico, sem fornecer ao aluno uma visão sólida dos fundamentos de aplicação geral. As raízes desse problema estão na própria natureza interdisciplinar da CIGeo. De forma direta, como ponto de convergência de áreas como Informática, Geografia, Planejamento Urbano, Engenharia, Estatística e Ciências do Ambiente, a CIGeo ainda não se consolidou como ciência independente. Para que isto aconteça, será preciso construir um sistema conceitual *ad-hoc*.

O primeiro passo para estabelecer as bases epistemológicas da CIGeo é identificar as fontes de contribuição teórica para reflexão. Esse objetivo torna-se mais palpável tendo-se em mira o conceito de *espaço geográfico*. Fundando-se nele, pode-se derivar um conjunto de conceitos que estabeleça a CIGeo como campo científico independente, com problema, objeto de estudo e metodologia definidos.

Esmiuçando-se mais o primeiro passo citado, é necessário *construir representações computacionais do espaço*. Assim, ao revisar as principais concepções da Geografia, na perspectiva da construção de sistemas de informação, contribui-se não apenas para a fundamentação teórica do *Geoprocessamento*, como ainda se busca inspiração para o projeto de novas gerações de SIG.

Esta pesquisa não tem por finalidade precípua fazer uma revisão das diferentes concepções de *espaço geográfico*. Pretende-se pincelar algumas noções e reflexões de intelectuais contemporâneos que estão preparando o terreno para acomodar esse novo rebento científico, e apresentar aos profissionais que dele se servirão um corpo validado de conhecimentos científicos, bem embasados em princípios éticos, sustentáculos da civilização.

Desse *corpus* e dos princípios éticos que todas as áreas do conhecimento humano devem possuir, é que se pode extrair conclusões sobre as capacidades e limitações das TIs de *geoprocessamento*; enfim, pode-se avaliar a sua utilidade social. Aí sim, a esteira comercial

---

[107] Nesta tese, denominada de CIGeo.

pode dar início ao seu movimento, lançando produtos que passaram pelo crivo da validação científica e pelo da sanção ética.

Os autores representativos foram selecionados de diferentes correntes da Geografia. No caso da *Geografia Regional*, cita-se HARTSHORNE (1936). Para a *Geografia Quantitativa* (no Brasil, também chamada de Teorética), citam-se HARVEY (1969) e CHORLEY (1967). No caso da *Geografia do Tempo*, cita-se HÄGERSTRAND (1967). Para a *Geografia Crítica* vêm os trabalhos de SANTOS (1996) e de HARVEY (1989). Neste tópico da revisão, ainda foram considerados os trabalhos de CHRISTOFOLETTI (1985), MORAES (1995) e CORRÊA (1995), também examinados na primeira parte da trilogia de CÂMARA (2002).

O subitem 1.2.3 e o Quadro 1.1 já resumiram esses enfoques. Outras considerações serão feitas a seguir, já que estão ligadas à natureza do problema de pesquisa

3.1.2.2. Geografia Idiográfica de Hartshorne e o *Geoprocessamento*

Em seu livro "Os Princípios e a Natureza da Geografia", R. Hartshorne procurou consolidar uma teoria para os estudos geográficos, baseada no conceito da *unicidade*.

Na sua visão, o objeto de estudo da Geografia seria "o estudo de fenômenos individuais" e a preocupação com o *único* na geografia não está limitada ao fenômeno, mas também se aplica aos relacionamentos entre os fenômenos".

O conceito de *unidade-de-área* é apresentado por Hartshorne como elemento básico de uma sistemática de estudos geográficos, denominada pelo autor de *estudos de variação de áreas*. Na visão de Hartshorne, uma *unidade-de-area* é uma parte do espaço geográfico, definida pelo pesquisador em relação ao objeto de estudo e da escala de trabalho, que apresenta características próprias.

As *unidades-de-área* seriam a base de um sistema de classificação e organização do espaço. Decompondo o espaço em *unidades-de-área*, o pesquisador poderá relacionar, para cada uma destas porções, as correspondentes características físicas e bióticas que a individualizam em relação a todas as demais propriedades do espaço. Hartshorne chamou este enfoque de *Geografia Idiográfica*.

A proposta de Hartshorne contribuiu para dar uma base metodológica para o uso do conceito de *unidades-de-área* em *Geoprocessamento*. A representação computacional correspondente a este conceito é o polígono, que delimita cada região de estudo e um conjunto de atributos, tipicamente armazenados num banco de dados relacional.

A geração atual de SIGs utiliza os conceitos de Hartshorne sobre delimitação de elementos espaciais homogêneos (*unidades-de-área*).

3.1.2.3. A Geografia Quantitativa e o *Geoprocessamento*

A base da *Geografia Quantitativa* (também chamada nos países anglófonos de *New Geography*) é a busca da aplicação do método hipótetico-dedutivo, que caracteriza as ciências naturais, aos estudos geográficos. Típico nessa perspectiva é o livro *Explanation in Geography*, de D. Harvey, que propõe uma aplicação dos paradigmas de generalização e refutação para os estudos geográficos amplamente utilizados por disciplinas como Física, Química e Biologia.

Ao criticar a falta de teorias explícitas na Geografia Idiográfica, os geógrafos desta escola passaram a utilizar teorias disponíveis em outras disciplinas científicas.

Pela perspectiva da Geografia Quantitativa, é preciso construir modelos a serem utilizados na análise dos sistemas geográficos. R.J. Chorley e P. Haggett argumentaram que esses modelos, construídos de forma teórica, devem ser verificados e validados com dados de campo e com base em técnicas estatísticas. Nesse contexto, o estudo dos padrões de distribuição espacial dos fenômenos (eventos pontuais, áreas e redes) passa a formar uma base para estudos quantitativos do espaço.

A Geografia Quantitativa também tem buscado subsídios na IA (RNA e Lógica *Fuzzy*).

Com a escola Quantitativa, os estudos geográficos passaram a incorporar o computador como ferramenta de análise. Nesse sentido, o aparecimento dos primeiros SIGs, em meados da década de 70, deu grande força à doutrina dessa escola. Ainda hoje, em países como os EUA, em que a Geografia Quantitativa é a visão dominante, os SIGs são apresentados como as ferramentas fundamentais para os estudos geográficos (CÂMARA, 2002)

Apesar do forte vínculo entre os conceitos da Geografia Quantitativa e o *Geoprocessamento*, apenas a partir de meados da década de 90 os SIGs passaram a dispor de representações computacionais adequadas à plena expressão dos conceitos dessa escola.

No estágio tecnológico atual do *Geoprocessamento*, ainda há restrições de representação dos fenômenos espaciais no computador de forma estática. A situação piora, quando um significativo conjunto de fenômenos espaciais, tais como escoamento de água da chuva, planejamento urbano e dispersão de sementes, entre outros, que são inerentemente dinâmicos, só contam com essas representações estáticas utilizadas em SIG, que não os capturam de forma adequada. Desse modo, um dos grandes desafios da CIGeo é o desenvolvimento de conceitos e técnicas que sejam capazes de representar adequadamente fenômenos dinâmicos, simulando-os não só no espaço mas também no tempo (BURROUGH, 1998).

O desafio de incorporação da Geografia Quantitativa aos SIGs ainda não foi plenamente vencido. Especialmente no caso de modelos para processos espaço-temporais, os SIGs

ainda se comportam mais como SICs, em virtude da natureza estática de suas representações computacionais.

3.1.2.4. A Geografia Crítica e o *Geoprocessamento*

A ênfase da Geografia Quantitativa no uso de grandezas mensuráveis para caracterização do espaço geográfico vem sendo objeto de fortes críticas nas últimas duas décadas. Essas críticas têm por argumento que, apesar dos resultados obtidos no estudo dos padrões espaciais, as técnicas da Geografia Quantitativa não conseguem explicar os processos sócio-econômicos subjacentes, nem capturar o componente das ações e intenções dos agentes sociais .

A visão dessa nova escola é ainda motivada por uma outra ideologia. Para os críticos mais extremados, a Geografia Quantitativa estaria comprometida com a visão associada à expansão do capitalismo e os muitos teóricos da Geografia Crítica tomam por base a filosofia marxista na construção de seus conceitos (CÂMARA, 2002)

Nesta tese, não se empreenderá uma incursão de cunho filosófico nessas polêmicas ideológicas, suscitadas por esta ou aquela escola. Importa considerar a relevância dos conceitos teóricos de "espaço", apresentados pelos proponentes da Geografia Crítica para o projeto de uma nova geração de SIGs. Nesse cenário, serão apreciados conceitos propostos por David Harvey, Manuel Castells e Milton Santos.

Milton Santos, em especial, foi um dos geógrafos mais empenhados em apresentar novos conceitos de espaço geográfico. Em seus trabalhos, o autor deu especial ênfase ao papel da tecnologia como vetor de mudanças da sociedade e condicionante da ocupação do espaço, no que denominou o *meio técnico-científico-informacional.* Apesar de enfatizar a contribuição da tecnologia para a Geografia, o autor não examinou em minúcia o problema do uso direto de ferramentas tecnológicas como SIGs em estudos geográficos. Mesmo assim, seus conceitos são extremamente relevantes para a definição de uma epistemologia da CIGeo, como se verá a seguir.

Em seu livro "Espaço e Método" (1985), M. Santos utilizou os conceitos de *forma*, *função*, *estrutura* e *processo* para descrever as relações que explicam a organização do espaço. A *forma* é o aspecto visível do objeto, referindo-se, ainda, ao seu arranjo, que passa a constituir um padrão espacial. A *função* constitui uma tarefa, atividade ou papel a ser desempenhado pelo objeto. A *estrutura* refere-se à maneira pela qual os objetos estão inter-relacionados entre si e não possui uma exterioridade imediata - ela é invisível, subjacente à forma, uma espécie de matriz, na qual a forma é gerada. O *processo* é uma estrutura em

seu movimento de transformação, ou seja, é uma ação que se realiza continuamente no tempo, alterando o estado do processo e tendo em vista um resultado qualquer.

A relevância desse conceito de "espaço" para a CIGeo é mais conceitual do que prática, visto que indica, essencialmente, limitações dos sistemas computacionais de representação da informação. Na atual geração de SIGs, pode-se caracterizar adequadamente a *forma* de organização do espaço, mas não a *função* de cada um de seus componentes. Pode-se, ainda, estabelecer qual a *estrutura* do espaço, ao se modelar a distribuição geográfica das variáveis em estudo, mas não se pode capturar, em toda a sua plenitude, a natureza dinâmica dos *processos* de constante transformação da natureza, em conseqüência das ações do homem.

Dessa maneira, as lacunas entre conceito e realidade revelam deficiências estruturais de todos os SIGs, no atual estágio do conhecimento. Para abrandar ou remover tais óbices, será preciso avançar muito na direção de técnicas de *Representação do Conhecimento e Inteligência Artificial* (SOWA, 2000), o que leva a considerações mais genéricas (e fora do escopo desta pesquisa) sobre as próprias limitações do computador como tecnologia de processamento da informação[108].

Em 1989, D. Harvey, em seu livro *"A Condição Pós-moderna"*, citado em CÂMARA (2002), trata das transformações passadas pela sociedade humana diante dos avanços das telecomunicações e da Informática, o que provocou o surgimento de novos conceitos sobre "espaço" e "tempo".

Valendo-se de observações de cunho econômico-financeiro, Harvey introduziu importantes noções sobre "proximidade", que acabaram por se incorporar às técnicas de Análise Espacial. A denominada "Lei de Tobler"[109].diz o seguinte:

"No mundo, todas as coisas se parecem; entretanto coisas mais próximas são mais parecidas que aquelas mais distantes."

O teor desta lei constitui parte fundamental da motricidade do PROFAX, ligando-se estreitamente aos conceitos de similaridade semântica, como se verá na revisão de literatura do subitem 3.2.

As conseqüências dessas idéias são de grandes proporções para a CIGeo, uma vez que a compressão do espaço-tempo que as TIs produziram na civilização atual subverte a lógica previsível de organização do espaço e estabelece um substancial desafio conceitual para sua representação computacional. Do ponto de vista da Análise Geográfica, os concei-

[108] O leitor interessado deve consultar PENROSE (1989) e SEARLE (1984), citados por CÂMARA (2002).
[109] De TOBLER (1979), *apud* CÂMARA (2002).

tos de Harvey significam que a forma tradicional de expressar as relações espaciais entre entidades geográficas (propriedades como *adjacência* e *distância euclidiana*) capturam apenas efeitos locais, e não permitem representar a dinâmica dos fenômenos sociais e econômicos do dia-a-dia.

Numa visão mais teórica do que prática, alguns autores referem-se a "espaços de geometria variável" [M. Castells *apud* CÂMARA (2002)], para designar a situação em que as articulações materiais entre os agentes econômicos e sociais ocorrem de forma muitas vezes independentes da contigüidade física. Esta situação envolve novos conceitos sobre "espaço", em que os fluxos passam a ser um componente essencial: fluxos de capital, fluxos de informação, fluxos de tecnologia, fluxos de interação organizacional, fluxos de imagens, sons e símbolos, todos elementos de cunho sociológico, que abrem um sem-número de questões, relacionadas à aplicabilidade geral da Lei de Tobler.

Sem abandonar as definições anteriores, mas buscando uma visão mais geral sobre os conceitos de espaço, Mílton Santos afirmava que "o espaço geográfico é um sistema de objetos e um sistema de ações". Seu objetivo, segundo CÂMARA (2002), era o de contrapor os elementos de *composição* do espaço (os *objetos geográficos*) aos condicionantes de *mudança* desse espaço (as *ações* humanas e os *processos* físicos ao longo do tempo). Numa formulação sintética, Santos enfatizava a necessidade de libertar o homem de visões estáticas do espaço, típicas do condicionamento secular dos mapas tradicionais. Para isso, ele inseriu a componente dos *processos variantes no tempo*, como parte essencial do espaço.

Do ponto de vista da **informação geográfica**, a noção de "sistemas de objetos" e "sistemas de ações" coloca-se num nível de abstração ainda maior que as formulações anteriores de M. Santos. Daí surgiram algumas **questões cruciais**: É possível realizar a transição desses conceitos abstratos para o âmbito de um sistema computacional? Quais as limitações da tradução das noções abstratas propostas para um SIG? (CÂMARA, 2002).

Numa primeira análise, a tradução do conceito de *sistema de objetos* e *sistemas de ações* para o ambiente computacional esbarra em outras três questões: Como modelar os *sistemas de objetos*? Como representar os *sistemas de ações*? Como expressar as interações entre os objetos e as ações? (CÂMARA, 2002).

Para representar os ***sistemas de objetos***, será preciso descrever cada um dos diferentes tipos de objetos componentes do espaço (ou da parcela do espaço em análise). É por isso que um dos avanços recentes na área de *Geoprocessamento* materializa-se no uso de **ontologias**.

Uma ontologia do mundo geográfico pode auxiliar produtores e usuários de SIGs do futuro em dois aspectos: 1°) A entender como diferentes comunidades compartilham informações; e 2°) A estabelecer correspondências e relações entre os diferentes domínios de entidades espaciais.

Genericamente, pode-se dizer que o uso de ontologias no campo de SIG é uma maneira de integrar técnicas de Representação do Conhecimento numa tecnologia com uma forte tradição geométrica e cartográfica.

Deve-se lembrar que, apesar do atraente emprego do novo conceito, o uso de ontologias em SIG enfrenta essencialmente os mesmos problemas das técnicas de Representação do Conhecimento (SOWA,2000). Esses problemas incluem a concepção de formalismos para armazenamento de informação e a tradução do conhecimento existente informalmente no domínio de aplicação para representações computacionais.

Vale lembrar, ainda, que a maior parte dos paradigmas atuais de Representação do Conhecimento são efetivamente estáticos, incapazes de modelar adequadamente a dimensão temporal e os relacionamentos dinâmicos e dependentes de contexto entre os objetos.

A representação dos *sistemas de ações* é ainda mais difícil num ambiente computacional. Sendo o computador uma ferramenta matemática e não analógica, a representação de processos depende fundamentalmente de modelagem numérica, freqüentemente realizada por equações funcionais. Cabe aqui distinguir dois grandes grupos de processos espaciais: os *modelos do meio físico* e os *modelos de processos sócio-econômicos* (que incluem os fenômenos urbanos). Esses grupos possuem variáveis e comportamentos diferenciados, que exigem diferentes tratamentos na fase de implementação.

Fenômenos físicos, tais como modelos hidrológicos e ecológicos, são exemplos de fenômenos com alto índice de variação do estado da superfície ao longo do tempo. Sua representação acurada depende da capacidade de derivar sistemas de equações que descrevam a variação espaço-temporal do fenômeno.

No caso de fenômenos sócio-econômicos, os processos têm uma complexidade muito maior, por envolver, além de fenômenos físicos, componentes de construção da realidade social.

Dessa forma, a aplicação do conceito de *sistemas de ações* à modelagem computacional de fenômenos socio-econômicos não pode ser reduzida à premissa funcionalista de que é possível derivar modelos matemáticos que descrevam o comportamento dos agentes sociais. Apesar disto, os autores consideram ser útil e válida a proposição de modelos que,

com crescente refinamento e inevitável *reducionismo*, possam simular parte do comportamento dos diferentes processos socio-econômico-ambientais (CÂMARA, 2002).

Em resumo, o conceito de Milton Santos de "espaço como sistemas de objetos e sistemas de ações" caracteriza um mundo em permanente transformação, com interações complexas entre seus componentes. Santos apresentou uma visão geral, que admite diferentes leituras e distintos processos de redução, necessários à captura desta noção num ambiente computacional. A riqueza inerente a esta noção está em deslocar a ênfase da análise do espaço e da *representação cartográfica* para a dimensão da *representação do conhecimento geográfico*. Afinal, como diz o próprio Milton Santos: "Geometrias não são geografias".[110]

3.1.3. Conclusão parcial da revisão de literatura da área de Cartografia

No subitem 3.1, foi elaborada uma revisão de literatura cujo objetivo foi alicerçar ainda mais o entendimento sobre as áreas recobertas pela CIGeo e pelas ciências cognitivas, indicando que esta sobreposição de problemas e de metodologias de pesquisa levará, inexoravelmente, a uma nova geração de SIGs.

Do artigo de BÄHR (1996) até a trilogia de CÂMARA (2002), a conclusão parcial que se tira desta parte da revisão é que SIGs que incorporem a semântica da IG em suas arquiteturas, efetivamente, ainda estão em fase de pesquisa. Comercialmente, ainda abundam os sistemas de natureza *idiográfica* e os quantitativos (1ª fase) do Quadro 1.1 (subitem 2.2.3).

Não se pôde avaliar até que ponto o mercado influi ou não na contenção de novidades, tis como: *ontologias*, *semântica*, SIGAIA, etc. e da sua incorporação às tecnologias de *geoprocessamento*. Por um lado, porém, foi possível verificar que em países desenvolvidos, valendo citar os EUA, Inglaterra e Alemanha, há verbas governamentais e privadas para financiar projetos-pilotos e para fomentar linhas de pesquisa de ponta, especialmente aquelas em que há interdisciplinaridade entre *geociências* e ciências cognitivas.

Talvez não haja qualquer tipo de influência intencional (de mercado ou de outro campo de poder) sobre o desenvolvimento de novos enfoques científicos para a IG, com as conseqüentes mudanças de plataformas das TIs de *geoprocessamento*. O que é bem provável que esteja acontecendo é que se vive numa região de limbo, entre um paradigma que se enfraquece diante de novas realidades e um novo paradigma, que se mostra mais adequado a lidar com essas realidades. Uma coisa é certa: as novas soluções tecnológicas não pode-

---

[110] EGENHOFER (2001b), no subitem 2.2.3, expressou: "...geometrias são importantes, mas não são essenciais..."

rão condenar todo um parque de sistemas (equipamentos e *softwares*) montado ao longo de mais de quatro décadas. A solução deverá basear-se no *constructo* da reutilização[111]. O que já foi dito sobre integração se ajusta ao caso (V. Figura 2.2).

A síntese do artigo de BÄHR (1996) consubstancia-se numa recomendação com a qual os pesquisadores de SIG da atualidade já estão preocupados. O teor da recomendação é o seguinte: "Não se pode tratar do conteúdo informativo de imagens, fotos digitais e outros produtos afins somente pelas *geometrias*". Há uma relação íntima entre a linguagem e o mundo-real (MR). A questão que se coloca é a de como encontrar uma teoria lingüística que seja adequada à recuperação da informação contida nesses documentos modernos e como tal teoria poderia dar suporte à interpretação da informação recuperada, criando uma teia de conceitos na cadeia de produção de conhecimento. Nesse ponto, não se pode mais ficar no domínio da Cartografia e das técnicas tradicionais de *geoprocessamento*. É hora de se socorrer nas ciências cognitivas.

O artigo de BÄHR (1996) parecia vaticinar a avalancha de novos conceitos que viria com a convergência de objetivos dos campos do conhecimento *geocientífico* e cognitivo, no final do século XX. O livro eletrônico (trilogia) de CÂMARA (2002), além de aprofundar os pressupostos e reflexões contidas naquele artigo de 1996, já foi suficiente para reunir condições de delinear o alcance e as limitações de uma nova ciência: a CIGeo.

Neste tópico da revisão de literatura, já antecedidos por considerações e definições preliminares dos dois primeiros capítulos, foram examinados diferentes conceitos sobre "espaço", originários de vertentes da Geografia e se buscou estabelecer que representações computacionais permitiriam a expressão desses conceitos no ambiente de SIG, que, não obstante saber-se que é uma visão reducionista e limitada, considerou-se que ela seja ainda muito útil, porque ajuda a compreender as diferenças entre os conceitos de espaço e os desafios ainda não resolvidos pela CIGeo.

O que se pode ainda concluir, parcialmente?

Em primeiro lugar, apesar dos significativos avanços nas duas últimas décadas, a tecnologia de SIG ainda está longe de dar suporte adequado às diferentes concepções de *espaço geográfico*. Atualmente, os SIGs oferecem ferramentas que permitem exprimir procedimentos lógicos e matemáticos sobre as variáveis *georreferenciadas*[112] com uma economia de expressão e uma profusão de repetição de pormenorizados processos automatizados de

[111] No caso de *hardware*: reaproveitamento.
[112] Neologismo designativo de informações ligadas a um sistema de posição espacial.

cruzamento[113] de dados para gerar informação útil, impossíveis de alcançar em análises tradicionais, mais próximas da linguagem humana.

Em segundo lugar, vem a outra face da moeda da modelagem formal mencionada acima. No patamar atual, a tecnologia de SIG resolveu apenas os problemas simples e adequados para a representação computacional do espaço. Os atuais sistemas são fortemente baseados numa *lógica cartográfica* do espaço, exigindo sempre a construção *de mapas computacionais*, tarefa sempre custosa e nem sempre adequada ao entendimento do problema em estudo. Ademais, mostrou-se que a Geografia Crítica tem uma importante contribuição para a CIGeo. Um dos principais méritos dessa vertente é o de sinalizar para um enfoque muito promissor sobre o espaço geográfico, pondo em relevo a noção do processo em contraposição à natureza estática dos SIGs de hoje.

É fundamental distinguir entre as capacidades da atual geração de SIGs e as limitações inerentes a qualquer representação computacional do espaço geográfico. Destarte, em que pese ser inexeqüível capturar num ambiente de *geoprocessamento* todas as dimensões de conceitos como *sistemas de objetos* e *sistemas de ações*, é importante buscar técnicas que propiciem aproximações dessas dimensões. Para tanto, será necessário adotar enfoques quantitativos, mas baseados em conceitos sobre *ontologias* e *representação do conhecimento*, sem perder de vista que os modelos inerentes a tais enfoques serão, no máximo, clones da realidade geográfica.

3.2. Revisão de literatura da área de ciências cognitivas

Esta parte é a que exigiu o maior esforço de pesquisa na revisão, visto que se trata da natureza mais íntima do problema de pesquisa e que sempre esteve presente, mas de forma tácita, no berçário de problemas *geocientíficos*, em particular, os cartográficos. Autores e pesquisadores de problemas lingüísticos já prognosticavam, na década de 60, que o estudo da linguagem codificada por formas multidimensionais de representação gráfica (mapas, imagens, gráficos) faziam parte de uma área interdisciplinar do conhecimento. Na verdade, muitos anos antes, no alvorecer do século passado, o próprio Ferdinand de Saussure já refletira sobre a semântica que fugia das amarras da dimensão linear do texto que corre pelas linhas do papel.

---

[113] Graças ao suporte oferecida por um SGBD comercial ou nativo do SIG.

Este subitem tem por objetivo revisar a literatura que trata da essência desse tipo de linguagem, que predomina num universo de comunicação multiforme, habitado por conceitos ligados a entidades espaciais e às diversas relações entre essas entidades.

É bom deixar bem claro que o objetivo supracitado não é e ainda nem poderia ser suficiente, por si mesmo, para apresentar uma revisão completa sobre o assunto, já que a literatura ainda é debutante e há muitas indefinições e conflitos que só o tempo poderá resolver. Contudo, nem por isso se deixará de pesquisar sobre o que já é de consenso ou sobre o que a experiência tem confirmado como evidência.

3.2.1. Considerações preliminares

No subitem 3.1.3, a síntese girou em torno da evolução e do impacto dos SIGs como um todo, dispensando certa atenção aos aspectos gerais de *software*. Neste tópico, os SIGs precisam ser observados não só pelo componente *software*. Há mais quatro outros componentes: *hardware*, *pessoal capacitado*, técnicas de *organização* e *dados*.

Para as ciências cognitivas, *software*, *hardware* e *pessoal* são domínios do conhecimento que assumem papéis protagonistas nas tentativas de engendrar teorias para explicar a criação de agentes inteligentes (AIs).

O componente de *hardware* (+ aplicativos específicos) de um SIG é uma parcela das TIs que também recebeu benefícios do progresso das telecomunicações e da Informática, sendo muito revelador, para delimitar o problema de pesquisa, discorrer sobre certos pormenores dessas TIs, conhecidas por *técnicas de gabinete*, em que os dados coletados de campo, pelas mais diversas formas, são transformados em mapas, que podem apresentar-se nos mais variados suportes (celulósico e digital, notadamente).

A Figura 3.1 é uma extrema simplificação das fases do mapeamento sistemático. Hoje em dia, seria uma tarefa difícil enumerar as várias formas de aquisição de dados (medições diretas do terreno por teodolitos, rastreadores de satélite, tomadas de fotografias aéreas, imagens de satélite, etc.), o que torna os instrumentos de tratamento desse amplo espectro de meios de aquisição de dados algo assemelhado a contos de ficção científica sobre robôs diligentes e dóceis, diante da flexibilidade e alto nível de interatividade a que estas *workstations* chegaram; para falar somente nas funções de aquisição e tratamento de dados!

A esquematização adotada na Figura 3.1 indica, do seu topo até a base, a seqüência de atividades que têm caracterizado o enfoque da escola da Geografia *Idiográfica* e as rápidas investidas na doutrina da escola da Geografia Quantitativa-1 (V. Quadro 1.1). Como se pode verificar, o mundo que é representável em refinados ambientes eletro-ópticos é um

conjunto de fenômenos de características estáticas, congelado num determinado momento de observação e cujo modelado é *restituído*[114] (reconstituído) pelas clássicas entidades gráficas euclidianas (pontos e retas) e por linguagens de programação algorítmicas.

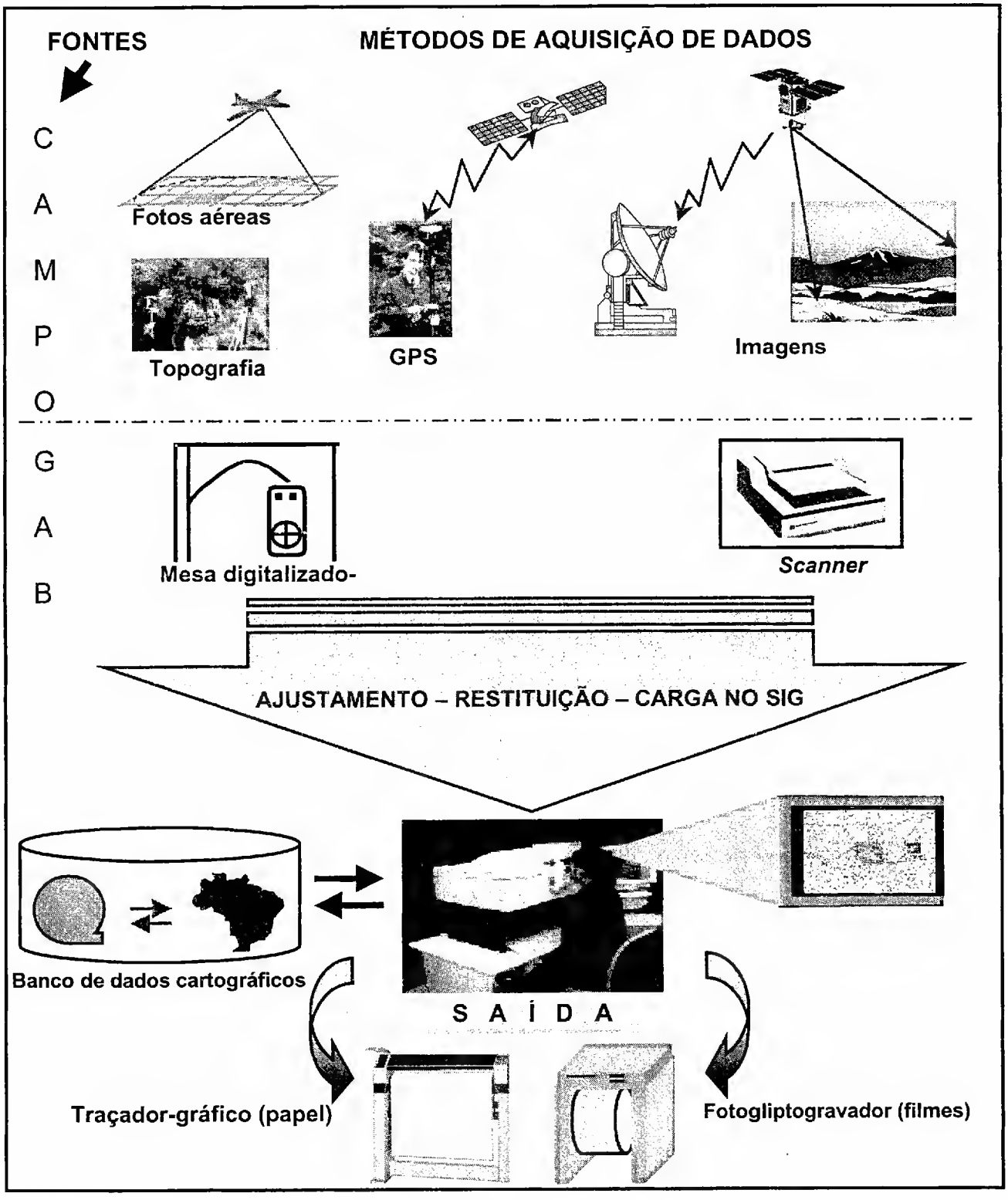

Figura 3.1: Ambiente típico de mapeamento com Cartografia Automatizada.

[114] V. *Fotogrametria* no glossário.

É fácil comparar a Figura 2.1 com a Figura 3.1 e notar a correlação entre as três etapas de transformação do MR para o nível simbólico, passando pelo nível iconográfico, de acordo com a primeira figura, e as etapas de aquisição, tratamento[115] e exibição de dados cartográficos, colhidos no terreno, na forma de mapas, de acordo com a segunda figura.

Os artistas, especialmente os escritores, sabem muito bem como ocorrem essas "transposições" do real (do fato) para a obra-de-arte (a representação). O fato é simplesmente o começo, o embrião da obra-prima para um poeta ou romancista[116]. Ao leitor de uma obra-de-arte como uma ficção, no entanto, não importa a exata comunicação do fato em si, percebido pelo escritor. O leitor deseja mergulhar no espaço de estimulação sensorial do autor, ou melhor, apreciar a sua criatividade e tentar usufruir do prazer (ou mesmo da contrariedade) que ele conseguiu incorporar ao resultado da transposição do real para as linhas escritas no papel.

Ainda dentro do exemplo de uma obra-de-arte escrita, o seu autor deve conservar uma atitude dualista diante do mundo factual, manancial de sua inspiração. Ao mesmo tempo que ele necessita manter-se em contacto direto com esse mundo factual inspirador, é pela mediação da linguagem (na forma escrita, no caso) que o autor conserva a necessária distância da realidade, sendo capaz de manuseá-la ao seu bel prazer, distorcendo-a ou alterando (ampliando ou reduzindo) o seu alcance semântico.

Produzir um mapa é conceber uma obra-de-arte, como vários autores já descreveram. Entender os processos de natureza cognitiva que presidem a construção dessas obras, para aproveitá-los em SIGs, será a tarefa dos futuros engenheiros ontológicos, que romperão, espera-se, as limitações que esses sistemas atualmente possuem, especialmente as limitações para representar fenômenos dinâmicos.

Mas foi em meados da década de 90 que já se notavam clamores acadêmicos por uma representação do MR que não fosse tão estática. As pretensões eram também dirigidas para formas alternativas de percepção do MR, e foi justamente sobre a fase de restituição fotogramétrica que convergiram as questões mais intrigantes do processo cartográfico de comunicação. Segundo a Figura 2.1, é nessa fase que o fato se transforma em fenômeno; noutras palavras, é nessa fase que os acidentes naturais e artificiais da superfície terrestre (ícones) transformam-se em imagens e mapas (índices e símbolos[117]).

---

[115] Cálculos de ajustamento, distribuição de erros e armazenamento de dados em bases de dados.
[116] Fonte: recente entrevista do escritor Salman Rushdie sobre o seu último romance: "O Chão a seus Pés".
[117] Segundo a visão triádica de Charles S. Peirce (Semiótica).

Como assinalou BÄHR (1996), a metodologia de produção de mapas, nos últimos 30 anos, passou por uma repentina mudança na forma instrumental de coleta, processamento e exibição de dados cartográficos. Numa rápida e instrutiva digressão desse autor, adapatada a seguir, é possível desvendar a ligação entre esse salto evolutivo nos métodos de aquisição e tratamento de dados cartográficos e os reclamos por uma tecnologia que "capture" elementos como tempo e comportamento do fenômeno espacial, como ocorre na Geografia Crítica.

Até meados da década de 70, pode-se dizer que o paradigma analógico ainda dominava os parques de produção cartográfica. Eram instrumentos eletro-óptico-mecânicos que restituíam o mundo-real geográfico em modelos de aparelho, calcados nos princípios da Geometria Projetiva e Euclidiana, havendo grande participação do operador humano neste processo (o que implica introdução de erros). Pouca ou nenhuma manipulação numérica era empregada, porque os procedimentos manuais de aparelho modelavam as estruturas físicas do mundo-real, substituindo-as, no papel, pelas abstrações geométricas (primitivas) de pontos e segmentos de reta. Era um pantógrafo[118] compondo o "esqueleto" do mapa (sistema artesanal de produção).

A partir de 1980, começaram a surgir soluções intermediárias entre a restituição analógica e a analítica (ou numérica). Os aparelhos ópticos (caros e muito precisos) foram aproveitados nessa fase como coletores de dados. O computador era utilizado para acelerar alguns procedimentos executados manualmente pelo operador, além de ser um depósito de dados coletados e processados.

O operador humano intervinha menos nessa fase (logo, menor possibilidade de erros). A chamada fase semi-analítica pode ser considerada, em termos evolucionários, como a analítica ou numérica, já que a totalmente analítica não foi tão comum (por causa dos custos). O importante é que os elementos portadores de informação da Geometria Projetiva da fase anterior (feixes de raios projetivos) foram substituídos por ternos[119] de coordenadas (pontos do espaço-objeto definidos por conjuntos numéricos).

No início da década de 90, começaram a surgir as chamadas estações digitais. Nelas, além do instrumental óptico de alta precisão, vinham embutidos programas que executavam todos os preparativos para medição, a medição e o posterior processamento dos dados. No entanto, diferentemente de conjuntos numéricos para representar a localização de pontos no espaço, esse é um modelo estocástico de operações sobre elementos estruturais da ima-

[118] Braços mecânicos munidos de uma ponta de grafite, ligados ao dispositivo principal de desenho (restituição).
[119] Coordenadas (x,y,z); "x" e "y", no plano; "z", profundidade ou altura. É um sistema tridimensional.

gem do fenômeno, transformando-o numa escala de tons coloridos ou de preto e branco, de base binária (*bits* e *bytes*).

O que ocorreu nessa evolução, em termos metodológicos e conceituais? Da analógica para a numérica, nada de relevante, cientificamente falando. Na analógica, a manipulação dos aparelhos para acompanhar o modelo tridimensional do terreno, vindo das fotos, é uma forma semântica de processamento que ocorre dentro da mente do operador, que vai percebendo visualmente os acidentes naturais e artificiais do terreno que é mapeado.

Na analítica, de certa forma, houve uma perda de conteúdo semântico no mapeamento, porque os processos mentais de percepção, envolvidos na fase analógica, são muito mais ricos que a proliferação de pontos que um computador efetua para cobrir o terreno mapeado da forma mais extensiva possível.

Em termos práticos, foi irrelevante a perda de conteúdo informativo da fase analógica para a analítica. Entretanto, da fase numérica para a digital, houve uma alteração significativa de instrumental e de sistema conceitual. Uma rede de pontos com coordenadas numéricas não expressa todo conteúdo informativo do fenômeno espacial em estudo como uma imagem digital, que contém uma escala de tons coloridos portadores de elementos informativos (atributos) e que podem ser combinados de formas só limitadas pela imaginação. O que representa esta combinação para produzir informação? É uma forma de linguagem, de comunicação. É um código que deve ser interpretado com recursos de mais alto nível de abstração do que a Geometria. É o domínio dos signos (Semiótica e Lingüística).

3.2.2. Revisão de literatura da área de Ciência da Computação

"Usuário e computador ainda estão longe de formar um casal perfeito na sociedade contemporânea. São inúmeros os problemas de comunicação entre as partes, causando mal-entendidos e frustrações pela incapacidade de ambos em 'exprimir' suas intenções." (BLASER, 2002)

3.2.2.1. Generalidades sobre a interação homem-máquina

O que BÄHR (1996) prognosticara em meados da década de 90 começou a tornar-se problema de pesquisa no início do fluente século.

Em BLASER *et al.* (2002), p.ex., o entendimento do processo de visualização é um estágio preambular para sistemas especialistas com capacidade de resolver problemas de SIG. Os autores abandonam as definições limitadas do termo e partem para um enfoque multidi-

mensional, em que a visualização está ligada não só à mera representação gráfica de objetos numa tela do monitor de vídeo de um computador, mas também à construção de uma *imagem mental* desses objetos. Esse enfoque implica integração entre o usuário e o sistema computacional, de tal sorte que ambos colimem um só objetivo: interação homem-máquina, em que a máquina (o sistema) tenha capacidade de capturar e interpretar estímulos oriundos de um usuário humano e que este não necessite de apurados conhecimentos tecnológicos para submeter as suas necessidades à máquina.

Percebe-se que esse é um típico caso de comunicação, envolvendo problemas de interface e linguagem. Como já se tocou nesse tema nos subitens 2.2.1 e 2.2.3 (SIS e CSIS), volta-se a confirmar que a Engenharia de *Software* tem pela frente um grande desafio.

O enfoque de BLASER (2002) é denominado de ***visualização preliminar*** e já se tornou uma linha de pesquisa nos círculos acadêmicos norte-americanos e europeus na área da CIGeo, sendo particularmente importante na área de SIG, já que os conceitos espaciais são de natureza muito complexa para serem tratados pelos sistemas tradicionais de interação, que desde a popularização das interfaces gráficas, materializadas nos computadores da Apple™ (o Macintosh®), no início da década de 80, pouco mudaram na interação homem-máquina, mais voltada para adquirir, tratar e analisar dados de base textual.

Esse caráter unidimensional[120] da interação homem-máquina também subverte a definição de visualização, colocando-a como etapa final no processo de produção da informação geográfica e não na sua gênese, como já foi discutido. Tal inversão de sentido acaba por deslocar o centro de gravidade das funcionalidades dos dispositivos físicos de entrada dos sistemas de informação (periféricos), mais comprometidos com textos, tabelas, enfim, produtos planares (bidimensionais) de bancos de dados relacionais. É por isso que ainda hoje a maioria das tarefas de aquisição de dados baseia-se na sua introdução por periféricos como o teclado ou o *mouse*, o que é satisfatório para a maior parte das aplicações triviais de automação de escritórios, mas não resolve problemas mais complexos da área do *Geoprocessamento*, tais como: cartas eletrônicas para monitoração de veículos, sistemas computacionais miniaturizados para rastreio de satélites (computadores em relógios de pulso!), sistemas de reconhecimento de voz para mapeamento, entre outros, em que os periféricos tradicionais de aquisição e de exibição não dão respostas satisfatórias.

Diante dos fatos apresentados até aqui, alguém poderia argumentar que as interfaces gráficas de hoje são muito amigáveis (*user-friendly*), ao alcance da maioria dos usuários lei-

---

[120] Mais focado nas limitações do sistema computacional do que nas reais necessidades do usuário.

gos em computação. Em que pese os populares ambientes de janelas, com seu menus recorrentes, comandos de arrastar-e-copiar e de apontar-e-desencadear, fornecidos pelo *mouse* e disponíveis na maior parte das aplicações comerciais, esse argumento não carrega de evidência o fato de que todas essas comodidades foram mudanças de método[121] e não de essência, i.e., a realidade multidimensional (incluindo o tempo) do fenômeno geográfico não foi contemplada. E é a forma de adaptar para um SIG a aquisição de dados orientada a texto, de característica seqüencial ou unidimensional, que faz a interação do usuário com esse sistema uma tarefa tediosa e pouco produtiva.

Este trabalho acompanha[122] uma linha de pesquisa semelhante à da *visualização preliminar*, preocupada com o aprimoramento de metodologias capazes de aumentar a interação homem-máquina, concebendo uma fase precursora ao tradicional módulo de aquisição de dados de um SIG, nada simples e cômodo de operar, sendo instrutivo apreciar alguns de seus aspectos.

Essa fase precursora (*visualização preliminar*) tem como condicionante aproximar-se mais do modo natural de se começar a formular um problema: **esboçando-o** (*sketching*), retratando-o por **gesticulação** (*gesturing*) ou **verbalmente** (*talking*). Essas três formas estão num nível muito alto de abstração, próximas do entendimento de usuários nada especializados nas minúcias de sistemas complexos de bancos de dados e de técnicas de *geoprocessamento*. Essa *fase-tampão* cria condições mais naturais e cômodas para a formulação gradual de um problema, evitando distorções posteriores, em fases mais formais de sua solução, já dentro do domínio do projeto ou da implementação.

Substituir as formas de interagir de humanos com máquinas, deslocando a importância dos papéis para o usuário, é substituir o teclado e o *mouse* por formas mais naturais de comunicação entre humanos: linguagem natural, gesticulação, esboço e - quem sabe? - até **pensamento**, se algum dia um sistema especialista conseguir captar o conteúdo informativo de uma consulta humana, interpretá-lo e produzir uma resposta adequada (nem precisaria ser definitivamente correta ou verdadeira, apenas adequada ao problema).

A Figura 3.2 mostra o deslocamento de importância dos papéis da máquina para o homem, no contexto da visualização. Em verde, a indicação é de pouca expressividade na interação homem-máquina e de franca utilização dessas interfaces nos dias de hoje; o amarelo é a situação recíproca (formas alternativas), passando por uma fase intermediária de verde e amarelo.

---

[121] Adequado para tarefas repetitivas, em que os problemas são mais simples e bem definidos do que os espaciais.
[122] Utiliza alguns desses conceitos, mas numa outra fase: a recuperação da informação geográfica.

Um oferecimento maior de opções de visualização é imperativo na conjuntura atual. Os desdobramentos advindos dos avanços da microeletrônica e da Internet põem em cheque as formas tradicionais de visualização. É um binômio quase antagônico entre os vastos *acervos de dados e informações*, espalhados por um sem-número de nós da rede mundial, e o aumento paralelo do *número de usuários*, ávidos por usarem tais recursos, mas sem tempo e (pior) capacitação técnica aprofundada sobre os recursos computacionais de que dispõem ou que estão distribuídos pelos nós da rede.

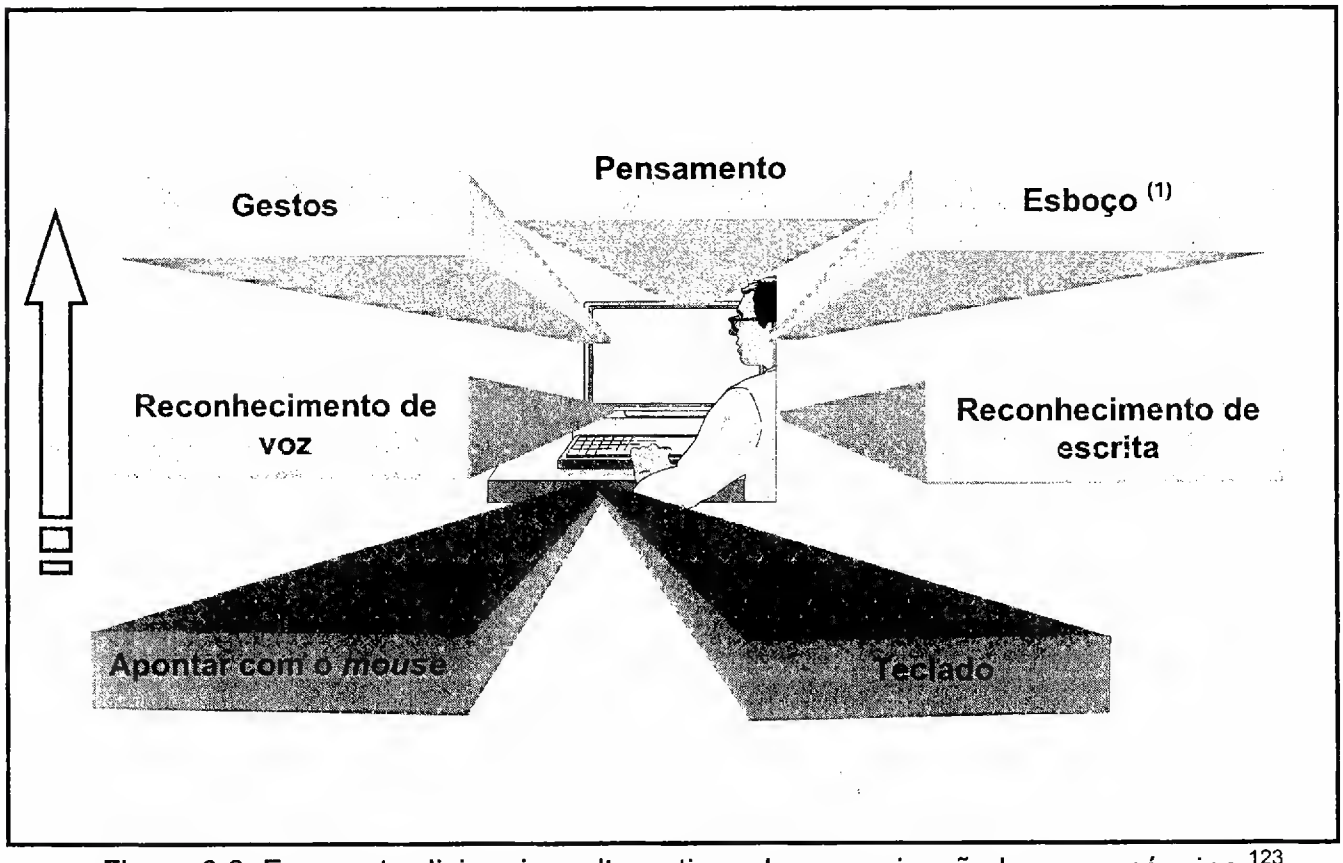

Figura 3.2: Formas tradicionais e alternativas de comunicação homem-máquina.[123]

Portanto, para envolver toda essa gama de usuários de *geotecnologias*, a correspondente gama de dispositivos e sistemas de visualização deve, igualmente, se estender e se conformar com a definição científica de ***visualização***: "Atividade que é capaz de produzir uma imagem mental de um fato observado e, ao mesmo tempo, de criar condições para que esta imagem mental possa materializar-se por meios gráficos" (BLASER, 2002).

Da definição anterior, é possível situar a visualização não só no domínio do sistema computacional, mas também no domínio cognitivo do usuário inteligente (humano ou não).

---

[123] Adaptada de BLASER (2002).

A *visualização preliminar* dedica especial atenção à fase crítica de formulação do problema. Sendo possível criar uma camada de *software* entre o usuário e as camadas que tradicionalmente já existem para implementar o processamento do problema, a obtenção e a análise dos resultados, como já discutido, também seria possível incrementar a gênese problema-solução com mais comodidade para o usuário e mais confiabilidade para os resultados obtidos na fase final de visualização gráfica [124], no domínio do sistema computacional.

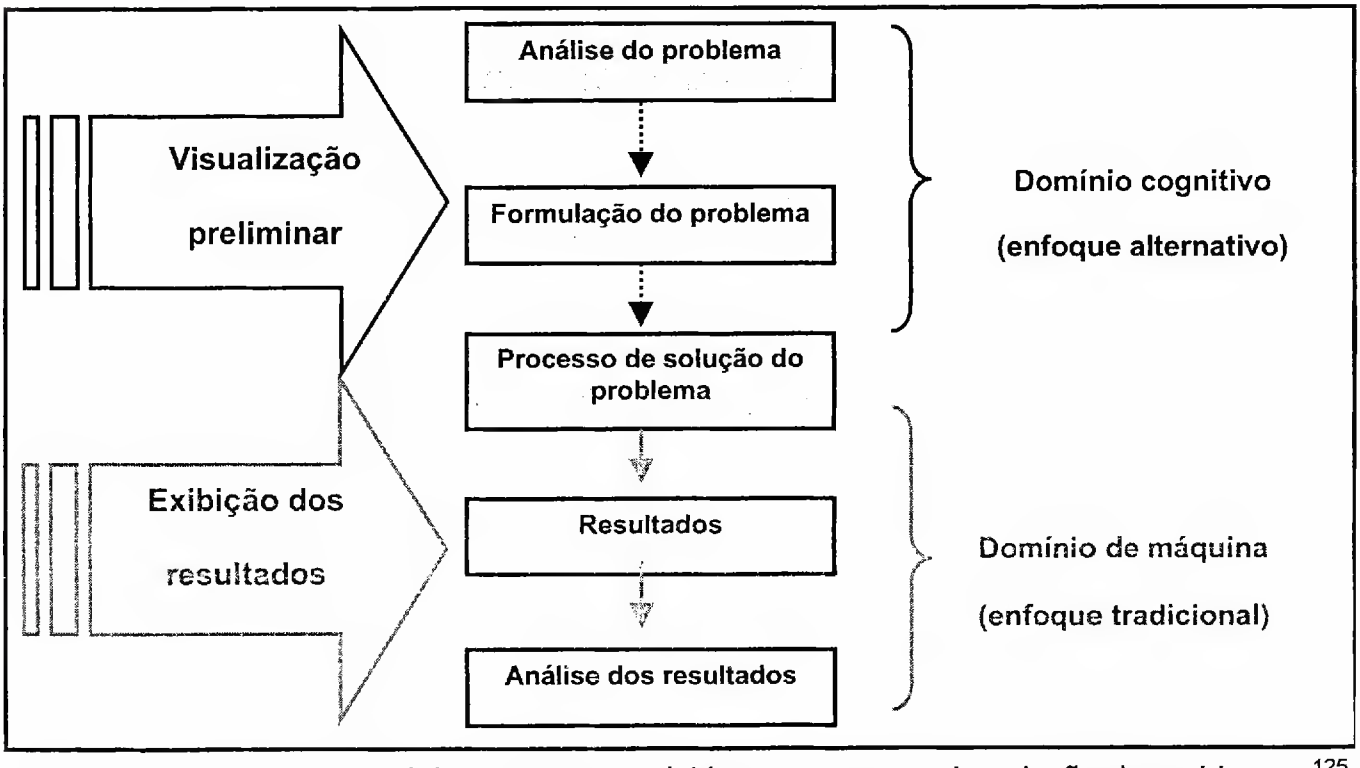

Figura 3.3: Seqüência geral dos passos envolvidos no processo de solução de problemas.[125]

A Figura 3.3 ilustra essa noção de melhoria do trabalho geral, em termos de qualidade, no encadeamento que vai da análise do problema até a análise dos resultados. A figura indica o enfoque tradicional da *exibição dos resultados* (na base) e o alternativo da *visualização preliminar* (no topo); este último enfoque pouparia o usuário de interagir com camadas mais rígidas e formais de computação. Esse aspecto do trabalho de BLASER (2002) também manifesta-se em RODRÍGUEZ (2000) e em todos os trabalhos que procuram por meios de promover os SIGs tradicionais para SIGs de características cognitivas.

---

[124] Neste ponto, o termo mais adequado seria *exibição dos resultados,* para diferenciar de *visualização preliminar.*
[125] Adaptada de BLASER (2002).

A definição de *visualização preliminar*, outrossim, coaduna-se com as etapas cognitivas do processo mental de modelagem da informação (atenção, indução e abstração)[126], um exercício quase sempre corriqueiro da faculdade intelectual humana, que segundo SAGAN (1977) e RUELLE (1993) vem sendo aprimorado desde o surgimento dos primeiros representantes do gênero *homo* (*homo erectus*, homem de Neandertal e *homo sapiens*, nesta ordem), há cerca de um milhão de anos atrás, na Época Pleistocênica, do Período Quaternário, da Era Cenozóica (LEINZ, 1980).

Já LOPES (1987) e ORTONY (1988) trouxeram à tona a chamada "metáfora científica", extrapolando as aplicações estilísticas tradicionais, em que se substitui o sentido natural de uma palavra por outro sentido, em virtude de uma comparação não enunciada. A metáfora sempre fora reduzida a uma matéria subsidiária da literatura. Nestas obras, contudo, os autores concordaram que a metáfora produz um "conhecimento revelado", que freqüentemente desafia os conhecimentos científicos construídos pela razão.

São vários os exemplos[127] que comprovam a força da metáfora como forma de incrementar a imaginação criadora do cientista ou do artista. É o resgate da força cognitiva desse instrumento de sondagem do raciocínio científico mais abstrato. A IA sai ganhando com a metáfora, visto que uma das contribuições oriundas da Lingüística Computacional deu origem a um dos ramos de pesquisa aplicada da IA: o PLN (Processamento de Linguagem[128] Natural).

Os conceitos dos trabalhos na região de limbo entre a Lingüística e a Computação não passaram despercebidos pelos cientistas da área geográfica. KUHN (1993), cientista da área de SIG, foi referenciado no artigo de BÄHR (1996) pelo seu trabalho intitulado *"Metaphors Create Theories of Users"*, cujo teor resume-se à confirmação do fato de que os seres humanos, nas suas constantes relações, utilizando-se primordialmente da língua natural, preferem também enriquecê-la ou complementá-la com outras formas de linguagem que não sejam as mais formais da Matemática e até mesmo as mais descritivas como a própria escrita. É intuitivo que as pessoas prefiram um esboço ou rascunho a um gráfico ou diagrama muito minuciosos; que prefiram a escrita na descrição de um fato a equações complexas da Álgebra; e que prefiram gestos ou a fala fluente à própria escrita.

Num contexto de aplicações espaciais, percebe-se como a *visualização preliminar* pode potencializar as formas de representação do pensamento dos seres humanos para que

---

[126] V. subitem 3.2.2.2.

[127] Guilherme Harvey, médico inglês do séc. XVII, num momento de *inspiração* (metafórica), associou os movimentos mecânicos de uma bomba aos da circulação sangüínea, estabelecendo as bases teóricas desse fenômeno fisiológico.

[128] A rigor seria Processamento de Língua Natural, mas aceitam-se ambas as acepções para PLN.

formulem os seus problemas. Essa fase prévia incorporada a um sistema especialista (SE) poderia ser capaz de orientar a intenção do usuário na solução do problema que imaginou, informando-o sobre a existência ou não de estruturas de dados na BC (base de conhecimentos) ou de regras no mecanismo de inferência do SE em questão, poupando-lhe tempo na desistência por uma infrutífera aplicação imaginada ou corrigindo-lhe os rumos, caso a aplicação fosse exeqüível.

BÁHR (1996) ressuscitou o paralelismo conceitual entre a modelagem lingüística e o processamento digital de imagens (PDI), já admitido por Saussure e inúmeros autores que o sucederam.

A semelhança de tratamento entre o signo verbal e o não-verbal reproduz-se até no acúmulo de imprecisões (polissemia) que o exercício da língua natural pela fala também repassa para a fase final de descrição simbólica do fenômeno geográfico (V. Figura 2.1), além das imprecisões trazidas das transformações anteriores. BÁHR (1996), portanto, reclamava da inexistência de uma teoria sobre essa propagação de erros em sistemas de representação do conhecimento (redes semânticas, p.ex.)

O próprio BÄHR (2000) aprofundou suas investigações de 1996 e acabou por realizar uma análise comparativa entre quatro formas de representação do conhecimento, muito utilizadas em SEs: as RNAs, as redes de Delaunay (RDs), as redes bayesianas (RBs) e as redes semânticas (RSs). Algumas conclusões deste trabalho serão vistas no subitem 3.2.2.2, por suas ligações com o embasamento metodológico adotado em RODRÍGUEZ (2000) e serão discutidos na concepção do modelo conceitual, especialmente no que diz respeito ao *contexto* para classificar imagens e recuperar informação geográfica de bases de dados.

O objetivo do trabalho de BÄHR (2000) foi resgatar a importância das *relações* tratadas nessas ferramentas de representação do conhecimento. Todas foram ordenadas segundo critérios que atribuem maior importância aos nós das redes (conceitos) ou aos arcos que ligam estes nós (as relações). Dessas comparações, BÄHR (2000) concluiu que o fenômeno espacial, assim como o verbal, é muito dependente do contexto, que pode ser adequadamente tratado pelos arcos das RSs, em que pese admitir e analisar combinações muito produtivas entre as quatro formas de representação citadas.

3.2.2.2. Revisão de literatura da área de Inteligência Artificial

"O cérebro humano é um processador iterativo, que quase nunca faz as coisas certas pela primeira vez. É um processo gradual de melhoramento. Portanto, se nós queremos aprender alguma coisa, não devemos tentar aprender tudo. Deve-se saber escolher o que adiar." (Gerald M. Weinberg)[129]

A IA é uma disciplina que ainda não está com o seu objeto consolidado, abarca problemas de muitas outras áreas do saber humano e tem sido incluída como disciplina da Ciência da Computação, na maioria dos estabelecimentos de ensino, especialmente nos programas de pós-graduação. Esta pesquisa seguiu este enfoque.

No entanto, alguns assuntos[130] já tratados nesta revisão foram incluídos fora da IA, noutros itens e subitens, por questões de relevância informativa e até de certa independência de seus conteúdos, mas, a rigor, deveriam estar inscritos no presente título.

O que se pretendeu nesta aparente cisão foi grupar na revisão de literatura de IA assuntos o mais estreitamente ligados ao objetivo desta pesquisa, especialmente os ligados à Lingüística Computacional e PLN, dos quais derivou o cerne desta tese: a similaridade semântica (SS).

O depoimento de vários semanticistas denuncia a inexistência de teorias semânticas viáveis a respeito de línguas naturais (LNs), mas isto não os intimida a deixar de continuar tentando e, pelo menos, delineando esboços dessas teorias. Todos admitem ser indispensável formular uma teoria que descreva as *significações* vinculadas pelas LNs.

O fato é que a *significação* está na raiz de todos os fenômenos dos quais o ser humano é participante, e nenhuma ciência poderá se constituir sem levar isto em conta [V. o trabalho de CRUZ (1994), no subitem 1.5.4].

Alguns desses semanticistas e terminólogos [(PINTO, 1977) e (FELBER, 1984, p. 160)] admitiram que essa lacuna teórica se deve a um certo descuido dos lingüistas que, desde Saussure, não se preocuparam o suficiente com os fundamentos da metodologia científica, exigidos pela ciência contemporânea, a fim de preservar a coerência e o controle na formação de conceitos, deixando documentados os conceitos adequados.

Nesse aspecto específico da formação de conceitos, não é raro quando pesquisadores de fora da Lingüística enveredam por caminhos da Semântica e acabam criando teorias in-

---

[129] *An introduction to general systems thinking* (1975).
[130] Visualização preliminar, p.ex.

coerentes, por causa de acepções equivocadas para termos como ***significado***, ***significação***, ***conotação*** ou ***intensão***, ***denotação*** e ***signo***, para citar alguns.

Apesar das restrições que PINTO (1977) colocou em relação ao *reducionismo* que os estudos[131] da Lingüística contemporânea está habituada a realizar, admitiu que os modelos formais da Lógica e da Matemática constituem um repertório bem definido e à disposição do pesquisador de Lingüística (Semântica), que não devem ser desprezados.

Dessa classe de estudos contemporâneos da Lingüística, uma espécie bem comum é a de procurar obter correspondência entre o fenômeno lingüístico e determinado sistema formal, o que se enquadra perfeitamente com a linha de pesquisa da CIGeo, ligada à modelagem do fenômeno da SS por modelos da Geometria Euclidiana e da Estatística. PINTO (1977) alertou para, simplesmente, não se assimilar o fenômeno ao modelo formal proposto, e tombar-se integralmente na cilada *reducionista*, que corrompe o objeto da ciência social de onde se originou o fenômeno.

Com o advento da IA e os problemas comuns com a Lingüística, surgiram linhas de pesquisa (PLN) e até disciplinas (Lingüística Computacional), em cujos *corpora* se justifica a aplicação e o concurso dos postulados, axiomas e até teoremas da Matemática em estudos sobre o ***significado***, visto que, segundo PINTO (1977), as definições deste termo e de seus correlatos originaram-se tanto da Matemática como da Lógica.

3.2.2.2.1. Tópicos relacionados aos aspectos semiológicos da informação geográfica

"Ora, em toda a terra havia apenas uma linguagem e uma só maneira de falar;"
(Gênesis, 11.1)

"Mapas são entendidos como veículos de comunicação"; este trecho do subitem 1.5.1 estabelece um vínculo forte com o presente tópico, ao chamar a atenção para uma das preocupações da IA: ***linguagens***; no caso em pauta, *linguagem cartográfica*.

É um grande desafio entender a linguagem cartográfica ao se relacioná-la com a IA. As formulações humanas sobre as relações espaciais são geralmente vagas, imprecisas, fortemente dependentes de contexto e influenciadas pelo ambiente de operação de um SIG.

---

[131] De caráter comparativo, especialmente.

Esse desafio aumenta, quando se leva em consideração a necessidade de implementar as diferentes formas de interação homem-máquina no momento da aquisição de dados, conforme ilustrado na Figura 3.2.

Ainda é insuficiente o conhecimento sobre as interações homem-máquina. O problema reside na dúvida sobre que grau aumentar a intervenção do sistema sobre a comunicação com o homem ou a situação recíproca. O balanço adequado nesses dois sentidos é o modelo sobre o qual cientistas da IA e de SIGs têm investido importantes esforços de pesquisa, que se valem da interdisciplinaridade com a Psicologia, sendo ainda necessário examinar alguns aspectos desta sobreposição de conhecimentos que o problema de pesquisa suscita.

A esta altura da revisão, é frutífero estudar algumas definições de conceitos que podem entrar na composição de um caso de avaliação de similaridade semântica de entidades espaciais, como o que se propõe para esta pesquisa. Tais definições vêm sobretudo da Lingüística[132], dentro dos limites estabelecidos pela Ciência da Computação (IA), e que determinarão a construção dos dois protótipos e do questionário que será aplicado aos indivíduos (metas operacionais). Essas definições ainda poderão servir de subsídio para fazer avançar para a fase de formulação de hipóteses e a fase do levantamento de variáveis em trabalhos futuros, bem como sistematizar essa literatura para leitores e profissionais que não lhe são afeitos, como os das engenharias. Todos esses elementos representam o quadro teórico que circunscreve o modelo conceitual proposto neste trabalho, cujo intuito é o de estender o entendimento do fenômeno da similaridade semântica.

A testificação da quase-imprescindibilidade dessa inclusão de conhecimentos, aparentemente alheios ao foco da presente pesquisa, veio de um processo crescente de revisão de literatura, que se expandiu, gradualmente, para a região de superposição entre estudos lingüísticos, *geocientíficos* e da Ciência da Computação.

Vários autores-cientistas do campo da Ciência da Computação e das *geociências* apresentaram e vêm apresentando, de forma cada vez mais profícua, artigos e mesmo livros que tratam de fenômenos lingüísticos (ambigüidade e metáforas, p.ex.) e mesmo psicológicos (aspectos cognitivos de percepção e elaboração do conhecimento), para chegar, de formas antes impensadas, ao fenômeno geográfico, que foi sempre tratado pela visão das engenharias tradicionais.

O incômodo inicial que o estudo de assuntos "estranhos" às ciências exatas causava, foi sendo dissipado quando se percebeu a estreita ligação que a *informação geográfica* (IG) possui com as idéias de *signo*, *símbolo*, *significação*, *simbolização*, entre outras; o que leva

a crer ser desastrosa[133] a empresa de se conceber um sistema de informação, na atualidade, que não leve em conta esse aporte conceitual multidisciplinar para ambientes complexos e dinâmicos de aquisição, análise e difusão de IG.

Desse modo, os conceitos pincelados das obras e artigos que se verão a seguir serão utilizados na seção de metodologia da tese, como aqui estão expostos ou como deverão ser coadunados com as necessidades daquela etapa do trabalho.

Outro fator importante para a inclusão dessas definições é garantir maior controle terminológico, quando se opera nessa região de limbo, evitando-se as indesejáveis: inadequação, obscuridade e confusão conceituais (V. *conceito inadequado* no glossário).

DUCROT (1972), ALSTON (1973, PINTO (1977), DAHLBERG (1978), SANTAELLA (1983), COUTO (1983), FELBER (1984), GUINCHAT (1994), CURRÁS (1995), GOMES (1998), COSTA (2002), CLUL (2002), FETI (2002) e DCCUT (2002) são autores e instituições de pesquisa que trabalharam com conceitos da Semiótica, da Semântica, da Lingüística e da Terminologia em suas obras e produções, manancial para um *corpus* de definições necessárias para construir uma rede semântica (taxinomia) que possa ser empregada num AI, como uma base de conhecimento.

Em quase todas as citações acima, uma preocupação comum pode ser decantada: o grande desafio para criar (esboçar) uma teoria semântica, porque se trata de conciliar aspectos até hoje aparentemente inconciliáveis entre os objetos das ciências sociais, de onde se originaram os problemas, e os modelos da Lógica e da Matemática, às vezes inconsistentes com a realidade complexa que os pesquisadores procuram sistematizar.

A. Culioli [*apud* PINTO (1977)] enumerou alguns comandos gerais de conciliação:

- O formalismo não deve ser uma simples *reescritura* de superfície (texto-livre), i.e., descrito em LN e sujeito à ambigüidade e ao subjetivismo do pesquisador;
- O pesquisador deve manter-se situado num nível de abstração[134], tal que o permita conduzir um estudo formal, buscando modelos adequados ao seu problema, controlando a terminologia e o uso de símbolos;
- É preciso retornar à superfície do discurso depois de se ter provado, no campo formal, uma hipótese ou pressuposto teórico. Esse retorno é a interpretação dos resultados.

Um outro problema surge na tentativa de formalização de uma teoria semântica aplicável às LNs. Trata-se da recorrência que o uso de uma metalinguagem para descrever o mo-

---

[132] Também incluídas a Semântica e a Terminologia.
[133] Como ocorreu com a recente (um pouco mais de 30 anos) história dos SIGs, mencionada nos itens preambulares.
[134] V. considerações sobre obras-de-arte no subitem 3.2.1 (visão dualista dos artistas).

delo escolhido pode provocar, por exemplo, na análise semântica[135], já que a descrição será vazada em LN, o que é inevitavelmente circular segundo T. C. Potts [*apud* PINTO (1977)]. O mesmo autor recomendou para que não sejam empregadas as mesmas expressões em ambas as instâncias da descrição (*definendum* e *definiens*), i.e., o que está sendo definido deve ser independente da expressão que define e vice-versa.

Como se vê, esses comandos gerais de Culioli e Potts cumprem a tarefa de cobrir as já mencionadas lacunas metodológicas de uma teoria semântica, e dão o pontapé inicial para um trabalho com desdobramentos no campo terminológico.

Antecedendo uma síntese sobre esta parte do trabalho, pode-se concluir que o que está por trás do fenômeno-alvo deste estudo – a similaridade semântica (SS) - é um porta-fólio de acepções, muitas das vezes conflitantes, sobre o ***significado***. Assim sendo, para estabelecer um meio comum de entendimento, é necessário descrever (sem citar controvérsias) as noções que mais se vinculam ao conceito de significado, que são: ***linguagem***, ***língua*** e ***fala***, de acordo com autores como DUCROT (1972), COUTO (1983) e CURRÁS (1995), que se serviram muito da obra de SAUSSURE (1975).

A ***linguagem*** é a ***faculdade*** humana absolutamente necessária para as finalidades de comunicação por intermédio da língua.

Essa definição genérica de linguagem não deu margem a discussões sobre a característica de a linguagem ser ou não inata aos seres humanos. O subitem 1.5.4 – As inteligências - já tratou disso *en passant*, no entanto, a definição (ou tentativa) anterior assumiu uma forma clássica e se refere à linguagem como um todo, porque nem sempre o ser humano se comunica por "sons". Muitas vezes ele se utiliza de sinais ou símbolos (mapas, Código Morse) e de imagens mentais como as metáforas. Por esse prisma, o conceito de *linguagem* ganha extensão, podendo-se considerá-lo como um conjunto de *símbolos convencionais*, *portadores de informação*, cujo principal fim está na comunicação entre os seres vivos.

Tomada em seu todo, a linguagem está a cavaleiro de diferentes domínios; ao mesmo tempo física, fisiológica e psíquica, ela pertence ao domínio individual e ao coletivo social. Não se sabe como inferir a sua unidade. A linguagem é heterogênea, difícil de ser definida.

A ***língua*** é um sistema arbitrário de signos fônicos e significativos, que satisfaz à necessidade de comunicação de uma comunidade humana. Como um todo, ela é uma entidade social, independente do indivíduo; a soma do conhecimento lingüístico armazenada no cérebro de todos os membros da coletividade; é, portanto, abstrata, um ***sistema***.

---

[135] Também denominada de análise *componencial* ou cálculo de predicados.

A língua não deve ser confundida com a linguagem; é somente uma parte determinada, essencial dela. É, ao mesmo tempo, um produto social da faculdade da linguagem e um conjunto de convenções necessárias, adotadas pelo corpo social para permitir o exercício dessa faculdade nos indivíduos.

A *língua* é a *parte social da linguagem*, exterior ao indivíduo, que, por si só, não pode nem criá-la nem modificá-la; ela não existe senão em virtude de uma espécie de contrato estabelecido entre os membros de uma comunidade. Além disso, a língua é homogênea, definível; constitui-se num sistema de signos em que, de essencial, só existe a união do *sentido* (significado ou conceito) e da *imagem acústica* (*significante*), sendo essas duas partes do signo psíquicas, mas nem por isso são abstrações intangíveis.

Uma acepção curiosa para *língua*, dada por PINTO (1977), vincula-a à noção de sistema, em que a significação de cada item constituinte depende da significação dos itens vizinhos, dentro de uma mesma zona de conhecimento.

Por essa visão de sistema, ao se acrescentar mais características distintivas ao termo *língua natural* (LN), gera-se uma espécie descendente na sua cadeia conceitual, a que PINTO (1977) chamou de *língua artificial* ou *formal*, construída racionalmente pelo homem.

Nessa classe de línguas formais, inserem-se as linguagens técnicas de programação – LTPs). Diferentemente das LNs, uma língua formal possui sintaxe e semântica bem definidas e independentes (o que não ocorre nas LNs), não conta com a propriedade de *consenso interpessoal* (como ocorre nas LNs) e pode ser manipulada pelo raciocínio humano (o que não é possível com as LNs).

Como acentuou PINTO (1977), para construir uma língua formal, pode-se estabelecer um programa de trabalho subdividido em duas etapas bem definidas e independentes uma da outra: uma é concernente às regras sintáticas gerais e a outra é concernente à metodologia de análise semântica, tendo em vista determinar a significação atribuída aos sinais, expressões e fórmulas (argumentos) escolhidas para montar o código dessa língua formal ou artificial (LTP, p.ex.).

As regras sintáticas gerais, acima citadas, pautam-se pelas seguintes diretrizes:

- Elaboração de uma lista dos sinais e dos seus tipos (morfologia);
- Relacionamento de regras de concatenação desses sinais, para ser possível formar expressões válidas na linguagem.

As regras de concatenação, por sua vez, subdividem-se em:

- Regras de formação, usadas para formar sinais simples ou palavras e, com base nestas, formar sentenças;

- Regras de transformação (ou de inferência), que permitem deduzir argumentos[136] de outros argumentos anteriores, sem levar em conta outras informações (contexto) que não sejam as fornecidas pela estrutura desses argumentos.

A montagem de ontologias não segue rígidos padrões de especificação, típicos das engenharias. Essas estruturas não constituem estruturas rigorosas de formalização do conhecimento, mas por se situarem numa região de limbo entre manifestações cognitivas do raciocínio e as representações formais adotadas pela Engenharia de *Software* para outros tipos de estruturas (técnicas de modelagem conceitual da UML™, p.ex.), costuma-se utilizar rotinas semelhantes às desta engenharia para explicitá-las, na falta de técnicas específicas.

Como se verificará, para a carga de uma dessas estruturas no PRONTO© (Capítulo 6), que é um produto de língua formal (*Java*™), ao mesmo tempo em que foram criadas definições de termos (parte semântica), aplicou-se um certo rigor de especificação na ordenação (parte sintática) desses termos, segundo as relações conceituais que se identificaram para eles, por meio de uma notação denominada *Backus-Naur Form* (BNF).

Nesse ponto da discussão sobre línguas formais, PINTO (1977) aproveitou para criticar as teorias transformacionistas (em que pese os ajustes) que pretendem tratar de forma independente a Semântica e a Sintaxe, no contexto das LNs, porque terminam quebrando um princípio fundamental e evidenciado da Lingüística: a inseparabilidade do significado e do *significante* do signo, como se verá a seguir. Uma das evidências vem da tradição gramatical, que nunca conseguiu definir as noções de Sintaxe independentemente das de significação (Semântica).

Como se verá a seguir, esse enfoque de língua formal de PINTO (1977) aproxima-se do de FELBER (1984), quando este caracterizou uma língua profissional.

A ***fala***, por seu turno, é a ***realização concreta***, individual, desse sistema - **a língua**. Esta não constitui uma função do falante, sendo um produto que um indivíduo registra passivamente. A ***fala***, ao contrário, é um ***ato individual*** de vontade (premeditação) e de inteligência, uma função do falante.

A Lógica até possui certa capacidade para analisar e criticar o *sistema da língua*, mas não pode fazer a mesma coisa com o *fenômeno da fala*.

Um enfoque mais objetivo sobre o conceito de *língua* é apresentado por FELBER (1984), que identifica esse conceito como a classe mais abrangente da seqüência descendente *língua geral – línguas profissionais e especiais – terminologia*.

---

[136] Conjunto de enunciados (premissas) relacionados entre si, que se transformam numa conclusão (SALMON, 1973).

A *língua geral* é o meio comum de comunicação que permeia uma sociedade, diferenciando-se numa ou noutra camada social por aspectos de ordem fonética, léxica ou semântica, que não prejudicam a comunicação básica.

A *língua de especialidade* ou especial é aquela que possui as seguintes características:

- Utilizada por grupos que estão à margem das camadas sociais de uso comum da língua geral;
- *Monofuncional*, i.e., a função comunicativa deve existir para permitir a comunicação do grupo que está unido em função de alguma tarefa de interesse comum, sem obrigatoriedade de compromissos éticos ou morais com a sociedade;
- Número limitado de usuários (grupo de marujos, bando de marginais, etc.);
- Geralmente o usuário aprende a língua de maneira voluntária, para poder ingressar e se comunicar no grupo;
- Essa língua não é imprescindível para a existência da sociedade.

A *língua profissional* tem as seguintes características:

- Também é *monofuncional*, porque é utilizada por um grupo de indivíduos que partilham uma determinada profissão ou um campo do saber;
- É obrigatório por todos os membros da profissão;
- Está sujeita à normalização, para evitar a ambigüidade na comunicação;
- É imprescindível para a existência do grupo que a usa e, de modo geral, de toda a sociedade que depende das profissões de seus indivíduos[137].

E a ***Terminologia***? Segundo FELBER (1984), está inclusa como uma espécie de língua profissional, em que o léxico, rigorosamente prescritivo e utilizado pelos indivíduos, é o seu campo de trabalho.

Como em diversas passagens deste trabalho há muitas referências a conceitos da Terminologia, é necessário situá-la no quadro referencial, para estudar mais adequadamente a SS. O enfoque de FELBER (1984) estabelece algumas definições para Terminologia, que, relembrando, a enquadram como uma língua profissional. Esta, por sua vez, é uma língua geral, que por sua vez é uma língua (natural – LN -, no âmbito da Lingüística).

FELBER (1984) oferece três definições para a Terminologia; a primeira privilegiando o *campo do saber* (ciência), a segunda, o estrito *campo terminológico*, e a terceira, o *produto terminográfico*[138]. Nessa ordem, estas são as três definições do autor para Terminologia:

- É o domínio do saber que trata dos conceitos e de suas representações;

[137] São evidentes os fatores de ordem sócio-lingüística presentes nas línguas de especialidade e nas profissionais.
[138] Chama-se Terminografia ao conjunto de métodos dedicados à produção de uma obra terminológica.

- É o conjunto de termos que representa o sistema de conceitos ligados a um determinado domínio do saber;
- É um acervo de publicações que contêm um conjunto de termos sobre um certo campo do saber.

FELBER (1984) elaborou essa classificação de domínios lingüísticos, para poder focalizar aqueles que mais valorizam o significado dos termos. A Semântica e a Terminologia têm muitos interesses em comum, por conseguinte, as suas unidades básicas de trabalho são o *conceito* e a sua materialização gráfica, o *termo*.

O ***termo*** é a designação de um conceito por meio de uma unidade lingüística definida numa Língua Profissional.

O ***conceito*** é uma unidade de pensamento, constituída de características (*intensão* e extensão), que podem ser aplicadas a uma entidade ou classe de entidades do MR (empírico ou imaginário).

Segundo MEDEIROS (1999), as restrições semânticas de um conceito são estabelecidas por sua *intensão* e extensão.

Segundo John Sowa [*apud* MEDEIROS (1999)], a *intensão* (ou compreensão) é o conjunto de todos os atributos ou propriedades que definem um conceito. A *extensão* é o conjunto de entidades às quais a *intensão* se aplica.

A *intensão* e a extensão variam em sentido inverso e possibilitam a hierarquização de um sistema de conceitos, base para a representação do raciocínio dedutivo (HUISMAN, 1976).

MEDEIROS (1999) acentuou a importância das noções de *intensão* e extensão dos conceitos para se estabelecer uma rede semântica (RS), que é uma forma de representação do conhecimento em IA (V. Figura 1.13). A prova disso é que a RS não foi escolhida apenas pela autora para construir o seu *corpus*, mas também a escolhida por RODRÍGUEZ (2000) e por inúmeros outros autores, não sendo exceção nesta tese, tampouco.

No subitem dedicado à compreensão das teorias do conceito (p.146), mais pormenores serão apresentados sobre *Terminologia*, *conceito* e *termo*.

Como já salientaram alguns semanticistas, não é nada trivial encetar um trabalho para criar uma teoria semântica, especialmente quando ainda ecoam grandes dúvidas quanto à delimitação do objeto dessa teoria com as seguintes questões:

- Onde acaba a língua e começa a fala?
- A Semântica é ou não uma disciplina lingüística?
- Onde termina a Sintaxe e começa a Semântica?

- E a Pragmática? onde colocá-la?

Parecem questões distantes do objetivo desta tese. Ledo engano, visto que o próprio título da obra inclui a palavra "semântica" e, como já enfatizado, o estudo do significado também é importante para uma teoria sobre SIGs que pretendem resistir ao tempo.

Já ficou bem claro que não se pode mais dissociar a IG do fenômeno da SS, que faz parte do objeto da Semântica e da Lingüística, mesmo diante das dúvidas sobre a independência - ou não - da primeira em relação à segunda. A evidência incontestável é que a Cartografia, a Lingüística e a Semântica partilham preocupações afins; e a IA, como última integrante do quarteto, parece ser a mediadora para produzir soluções, mesmo que ainda em domínios limitados.

Das questões acima, ainda surge o campo da Pragmática, com implicações em várias teorias do conceito, especialmente quando ocorre informação contextual. O que ela e suas congêneres - a Sintaxe e a Semântica - significam ficará mais claro, situando-as umas em relação às outras.

R. Carnap [*apud* PINTO (1977, p.19)] ofereceu uma descrição sistêmica para o trio de disciplinas lingüísticas, ao estabelecer *parâmetros funcionais* como a *relação* entre as *expressões* e as *entidades* designadas e o *uso* que se faz das expressões, resultando:

- Faz-se **Sintaxe**, quando são analisadas as relações entre expressões, abstraindo-se as entidades designadas e os usuários;
- Faz-se **Semântica**, quando são abstraídas as relações entre as expressões e os seus usuários e o foco é sobre as entidades designadas;
- Faz-se **Pragmática**, quando são abstraídas as relações entre as expressões e as entidades designadas e o foco é sobre os seus usuários.

Pelo prisma filosófico, a Semântica às vezes é colocada como o ramo da Lingüística que se ocupa dos significados das palavras [Michel Bréal, *apud* MORA (1994)]. Outras vezes é colocada como ciência independente, que estuda as diversas relações das palavras com os objetos por elas designados ou, como defende Américo Castro [*apud* MORA (1994)], a ciência que estuda as mudanças dos significados das palavras, visão compartilhada por PINTO (1977), que a complementou com a descrição *ut retro* de R. Carnap.

Para os filósofos, em geral, a *Semântica Lingüística* (descritiva para alguns semanticistas) é uma ciência empírica, em que a *indução* é o método por ela usado para a formulação de suas leis. Para os mesmos filósofos, também existe uma Semântica Pura, de base formal, totalmente analítica e sem conteúdo factual, em contraposição à Semântica Lingüística.

Esses filósofos imaginam as três disciplinas dispostas numa pirâmide, em cuja base ficaria a Pragmática; na porção central, a Semântica; e, no cume, a Sintaxe. As variáveis que orientariam a disposição do trio de disciplinas lingüísticas nessa pirâmide seriam a *formalidade* e o *nível de abstração*. Quanto mais para o cume, aumentaria em intensidade a *formalidade* e *o nível de abstração*; quer dizer, a Sintaxe, tratando de sistemas de signos não interpretados, seria a mais abstrata e formal; a Semântica, tratando de sistemas de signos interpretados, a medianamente abstrata e formal; e a Pragmática, tratando do uso que indivíduos inteligentes fazem desses sistemas de signos, seria a menos formal e a mais concreta das três.

Para sistematizar esta revisão, colimando sempre o objetivo geral da tese, foram abertos mais dois subitens neste tópico, para enriquecer o marco teórico do trabalho.

O primeiro subitem é dedicado a uma sinopse sobre as preocupações da Filosofia da Linguagem com relação ao ***significado***. Serão folheadas as três[139] linhas de pensamento filosófico que mais influenciaram as teorias sobre recuperação da informação e similaridade semântica (SS); entre os que as conceberam, citam-se: o filósofo e lógico alemão, Gottlob ***Frege*** (1848 – 1925), com a ***Teoria do Sentido***; o filósofo inglês, Bertrand ***Russell*** (1872 – 1970), com a sua ***Teoria do Atomismo Lógico***; e, finalmente, o filósofo austríaco, Luís (Ludwig) ***Wittgenstein*** (1889 – 1951), com a sua ***Teoria Pictorial da Frase*** (COSTA, 2002).

Os dois últimos influenciaram-se mutuamente e receberam muita influência do primeiro, que, por questões de ordem cronológica, foi o pioneiro e não manteve tanta interação com seus outros colegas, apesar da sobreposição dos seus períodos de vida e de produção científica.

O segundo subitem é dedicado às teorias do conceito, das quais muitos sistemas de informação (SIs) têm recebido contribuições de ordem prática. Além dos aspectos teóricos que serão vistos, muitos dos requisitos levantados para a implementação dos protótipos emanaram das diretrizes e normas que foram pinceladas do referencial teórico sobre noções de ***signo***, ***significante***, ***significado***, ***significação***, ***conceito*** (suas relações), ***denotação***, ***conotação*** e ***definição***.

Os *constructos* citados acima são parte de uma extensa teoria científica (de natureza multidisciplinar) e às vezes de acaloradas e controvertidas discussões que extravasam o campo científico e invadem as reflexões de ordem filosófica. Será de bom alvitre pincelar algumas definições e considerações que exploram o que de mais consensual existe sobre

---

[139] ALSTON (1973, p.27), por outro enfoque, cita a referencial, a *ideacional* e a *comportamental.*

essas noções fundamentais, para se entender o fenômeno da informação geográfica nas suas manifestações iconográficas e simbólicas (Figura 2.1), fulcro desta pesquisa.

Um registro preliminar faz-se necessário, antes de se passar aos dois subitens: "Esta tese não operará com entidades frasais, mas tão-somente com termos (nomes) que denotam os elementos espaciais (geográficos) do mundo-real".

De certa forma, trabalhar com estes elementos constitutivos das frases (os termos) têm sido a chave do sucesso dos SIs, dentro da lógica da IA de rejeitar problemas NP (V. glossário), naturalmente gerados se o elemento de estudo for a frase.

É de bom alvitre que os desafios para o entendimento de entidades tão complexas como as frases, matrizes dos enunciados e proposições que povoam o pensamento humano, fiquem por conta das reflexões filosóficas e teorias delas originadas no campo da linguagem, porque, dos grandes esforços empreendidos na busca da verdade e na tentativa de explicar racionalmente a frase, não são raros os malogros e muitas incoerências, quando é necessária a confrontação das hipóteses com a realidade empírica. Mesmo assim, trabalhos como o de MEDEIROS (1999) estão na vanguarda de estabelecer as capacidades e os limites de SIs que necessitem operar com texto-livre (frases). Os que estão na linha de RODRÍGUEZ (2000), como é o caso desta tese, operam com termos.

Apontamentos de Filosofia da Linguagem: esta concessão de espaço que se faz à Filosofia da Linguagem foi uma necessidade sentida por um pesquisador oriundo das ciências exatas, que avançou os limites de sua formação profissional e acadêmica, particularmente num estudo exploratório como este.

Para demonstrar que esta inserção de assunto de cunho filosófico não foi um desperdício e nem uma falha de revisão de literatura, optou-se por ficar na companhia segura do que exortou o biólogo e filósofo francês Jean Rostand [*apud* HUISMAN (1976, p.7)].

O teor da exortação de Rostand se prende à sua percepção sobre a crescente e inevitável cooperação entre o cientista e o filósofo na produção de conhecimento científico, uma vez que o primeiro não deve se constranger em aceitar os esclarecimentos do segundo, quando, em suas investigações teóricas, "deixando o seu próprio terreno", estabelecer questões de ordem geral que confinam com a Filosofia, o que ocorre amiúde em estudos exploratórios.

Evidentemente, nessa interação, o filósofo não sai perdendo, porque acaba por exercitar seu poder de reflexão e por cumprir um de seus mais relevantes papéis: criticar o valor do conhecimento em novos campos que a ciência está sempre desvendando.

Nessa altura da revisão, já é possível entrar na ***Filosofia da Linguagem***, para a qual COSTA (2002) cita dois ***enfoques*** sobre o ***significado***: o *geral*, preocupado com a crítica da linguagem, seus problemas e metodologia e, o que mais interessa a este estudo, o *estrito*, preocupado com a análise da linguagem, sua natureza e suas funcionalidades.

Tanto um como o outro enfoque estão na pauta de interesse de duas espécies de Filosofia da Linguagem: a *ideal* e a *ordinária*.

A *Filosofia da Linguagem Ideal* é influenciada pela Lógica de Primeira Ordem (LPO), baseada essencialmente nos trabalhos de Frege.

O método de cálculo de predicados desenvolvido por Frege foi (e ainda é) muito utilizado por lingüistas que se interessam pela formalização da linguagem e pela IA – os lingüistas computacionais. Esses estudiosos procuram formas de revelar, com o auxílio da computação, a verdadeira estrutura lógica que está por trás das sentenças da língua natural, que às vezes é muito diversa da sua estrutura aparente. O precursor desses estudos, nesses moldes, foi Frege, que lançou os fundamentos de uma teoria esclarecedora sobre o significado (COSTA, 2002).

A *Filosofia da Linguagem Ordinária*, por seu turno, toma como modelo a linguagem quotidiana, do senso comum, tentando investigar a sua estrutura funcional.

A IA e, em particular, a linha de pesquisa sobre SS, dentro da CIGeo, sustentada no trabalho de RODRÍGUEZ (2000), tiram proveito das preocupações dessas duas espécies de Filosofia da Linguagem. Prova disso são os trabalhos de EGENHOFER (1995) sobre a língua profissional geográfica, para orientar o desenvolvimento de sistemas de informações geográficas apoiados pela IA (SIGAIAs).

A ***Teoria do Sentido*** (significado) de ***Frege*** baseia-se no princípio da expressividade do pensamento ou da importância que um enunciado ou proposição têm como portadores da verdade (V) ou falsidade (F).

Para Frege, o significado ou sentido (em alemão, Sinn) é o portador da verdade de um enunciado. "Il pleut", "It is raining" ou "Está chovendo", apesar de serem frases totalmente distintas na forma, compartilham de algo comum, "querem dizer" a mesma coisa, são todas frases que possuem um sentido comum, um portador *fregeano* da verdade que está por trás de suas formas (escrita ou fonética): o ***significado*** (COSTA, 2002).

Para Frege, por conseguinte, **portar significado ⇒ comunicar**, i.e., para existir comunicação (condição necessária: sem ela o fenômeno não se produz) é preciso que o enunciado seja portador de significado, que veicule informação (condição suficiente: presente, produz inevitavelmente o fenômeno).

Foi refletindo sobre as frases de identidade, que Frege deduziu o significado (o 3º elemento, transcendendo o objeto e o nome que se refere ao objeto). Por exemplo[140], na frase: "A estrela da manhã é a estrela da tarde" há informação sendo veiculada pelo significado, o que não ocorreria na frase: "A estrela da manhã é a estrela da manhã".

Na primeira frase, os nomes "estrela da manhã" e "estrela da tarde" se referem ao mesmo objeto, o planeta Vênus. A forma como Vênus (o objeto) é referido (como se tem acesso à observação do fenômeno) é que veicula informação, que estende o conhecimento sobre o fenômeno. Vênus, como estrela da manhã, possui um comportamento que o distingue da forma como é chamado de "estrela da tarde". No primeiro caso, Vênus é o astro mais brilhante antes do sol nascer. No segundo caso, o mais brilhante antes do crepúsculo. Na identidade dessa frase, o que causou a modificação entre o que se quer dizer com cada sinal é o que Frege chamou de sentido ou significado.

Na segunda frase, a redundância (ambigüidade) é vazia de conteúdo informativo, porquanto "quer dizer" o que o interlocutor no processo de comunicação já conhece. E como só há fluxo de informação quando "se diz algo" sobre uma coisa que se desconhece, esta frase não possui o mesmo elemento portador de informação que a primeira: o significado.

Vale mencionar que a Teoria do Conceito da Terminologia científica de Würster acabou incorporando os conceitos adequados de todas as três correntes da Filosofia da Linguagem - de Frege, de Russell e de Wittgenstein. Uma prova cabal dessa verificação é o esquema da Figura 3.7, que é a representação do significado como o 3º elemento, o amálgama dos três vértices do triângulo.

Mas a expressão do pensamento por frases, segundo Frege, também possui uma referência, além do componente portador da verdade (ou falsidade). E é aí que a sua teoria cai no implausível, apesar dos "subprodutos" muito promissores que forneceu à ciência, como a LPO e a *Teoria do Contexto* (MILLER, 1991), intensamente explorada na tese de RODRÍGUEZ (2000) e, de certa forma, no PRONTO©, como se verá no Capítulo 6.

A *Teoria do Contexto* vem do *princípio leibniziano da intersubstituibilidade*, ou seja, é possível que uma expressão seja substituída por outra na frase, sem que o valor-verdade da frase se altere, sempre que a expressão que substitui se refira à mesma coisa que a expressão substituída [(MILLER, 1991) e (COSTA, 2002)].

Frege introduziu inconsistência à sua teoria mais abrangente sobre o significado, quando admitiu que a referência de uma frase podia ser traduzida por valores dicotômicos (ou

---

[140] Exemplo retirado de COSTA (2002), com as adaptações de responsabilidade do autor desta tese.

booleanos) como V ou F. A admissão disso pode até ser logicamente plausível em alguns casos, mas como a referência de um enunciado ou mesmo de um vocábulo (nome) requer ligação (denotação) com o mundo da realidade empírica, torna-se implausível fazer referências entre nomes e entidades do mundo-real apenas por valores como V ou F.

Outro passo importante para uma teoria do significado foi o trabalho de ***Russell***, lançando fortes alicerces para uma teoria das descrições, a que denominou de ***Teoria do Atomismo Lógico***.

Russell distinguiu duas espécies de ***conhecimento:*** o ***familiar***, muito ligado à noção filosófica de ***intuição*** e o ***descritivo***, muito ligado à noção filosófica de ***raciocínio***, sendo conveniente descrevê-los com base nessas noções, já vistas no subitem 1.5.4. O Quadro 3.1 confronta esses dois tipos de conhecimento, sintetizando as suas características.

Quadro 3.1: Conhecimento familiar e descritivo

Intuição (Conhecimento familiar ou imediato)	**Raciocínio** (Conhecimento descritivo)
Imediato (sensível)	**Mediado por conceitos**
Inexprimível	**Comunicável (linguagem)**
Singular	**Complexo**
Isolado	**Relações entre sinais (sons, vocábulos)**
Intensivo, compreensivo ou de conteúdo	**Formal**

Como se percebe do quadro comparativo, apesar de aparentemente antagônicas as características do conhecimento familiar (intuição) e as do descritivo de Russell, ambas entram num processo que flui seqüencialmente na obtenção de conhecimento, ou seja, são duas faces da mesma moeda. A intuição fornece a matéria-prima do conhecimento e, daí, vem o esforço racional do conhecimento descritivo, processando os dados fornecidos pela primeira. Não haveria conhecimento científico sem o concurso de ambos.

O Atomismo Lógico de Russell recebeu influência de Frege e, mais ainda, de Wittgenstein, especialmente pela alternativa ao *valor-verdade* do primeiro que o segundo estabeleceu, focando o seu interesse não no mundo ideal da verdade ou falsidade de um enunciado, mas na referência de uma frase no fato que ela possivelmente designaria, entendendo-se *fato* como um complexo de entidades e suas relações existentes no mundo extensional, sujeito à experiência cognitiva humana (RUDIO, 1978).

No Atomismo Lógico, é muito importante a concepção da ligação linguagem-mundo, em que, se todas as sentenças da linguagem humana forem devidamente analisadas, seria possível formar um conjunto de signos fundamentais (atômicos), representativos de partes da realidade do mundo.

Assim, para Russell, os verdadeiros nomes tinham significado pelo seu poder de denotação, isto é, pela capacidade que possuem de indicar ou de apontar entidades com as quais o homem mantenha direta familiaridade. Ao estender essa noção para elementos frasais, Russell começou a produzir inconsistências no Atomismo Lógico.

Encerrando esta visão geral da Filosofia da Linguagem, vem à tona a obra de referência de ***Wittgenstein***, *Tractatus Logico-Philosophicus*.

Apesar de bem situado na Filosofia da Linguagem Ideal, quando escreveu o ***Tractatus***, o filósofo austríaco, numa segunda fase da sua produção intelectual, acabou por pender para a Filosofia da Linguagem Ordinária, pela necessidade que teve de lidar com aspectos mais práticos da vida (COSTA, 2002).

O objetivo do *Tractatus* foi o de explicar a natureza representativa (Wittgenstein denominou-a de *pictorial*) ou factual da linguagem, i.e., como, por meio dela, o homem seria capaz de compreender o mundo. A análise da linguagem seria o seu instrumento básico nessa realização. Estes são os traços marcantes da ***Teoria Pictorial da Frase*** de **Wittgenstein** (COSTA, 2002).

A função de representação da linguagem foi aproveitada por muitas disciplinas, como a IA, que estudam fenômenos como o da SS.

O interesse dos estudiosos da IA pela teoria de Wittgenstein deve-se à sua natureza mais resistente às críticas e por ser a confluência da maior parte dos conceitos adequados das teorias de Frege e de Russell., apesar de também possuir limitações.

Wittgenstein admitia que os enunciados em língua natural poderiam ser decompostos em combinações de frases elementares (**modelos** de realidades elementares), que, por sua vez, seriam combinações de elementos ainda mais simples: os nomes dos objetos ou termos: "Algo rígido, imutável, indivisível - as pedras de construção do mundo!", num sentido metafórico atribuído por Wittgenstein.

O *constucto* **modelo**, para Wittgenstein, conforme citado acima, é um meio (visual, sonoro; documentário, enfim) capaz de representar a realidade, uma situação externa, devendo existir algo compartilhado entre ambos, modelo e realidade, ao que Wittgenstein chamou de forma lógica ou ***função afigurante*** (seria a denotação terminológica).

O interessante da teoria de Wittgenstein são os pontos de contacto com outras teorias, em especial, a do conceito, adotada pela Terminologia.

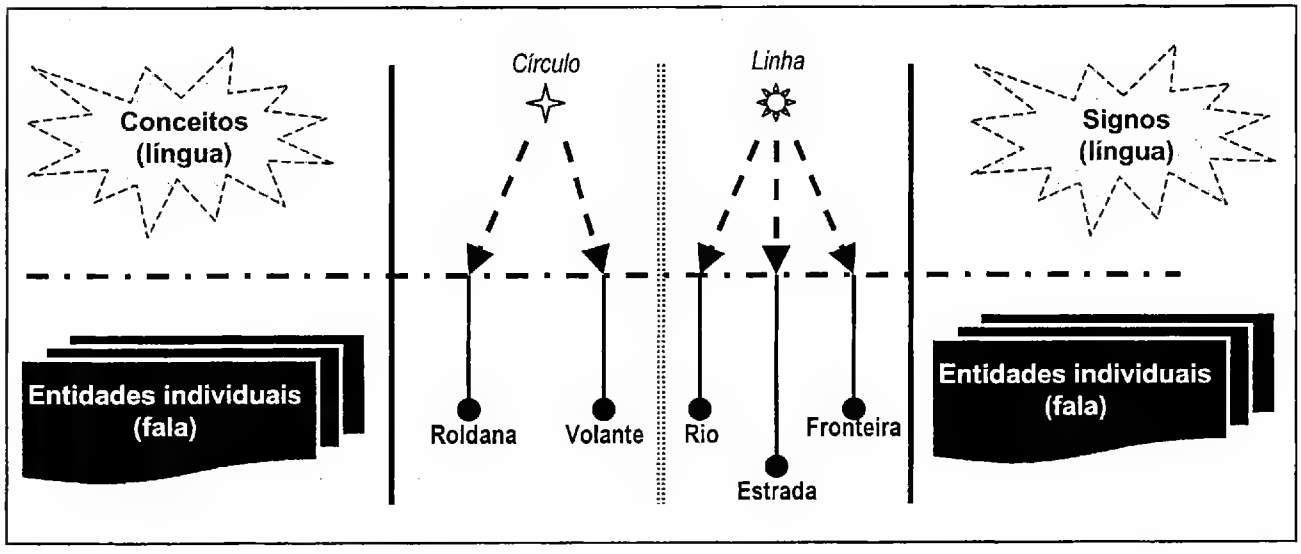

Figura 3.4: Sistema conceitual de FELBER (1984).

Da revisão de literatura, percebeu-se, p.ex., que o terminólogo FELBER (1984), sem dúvida influenciado pelo seu autor predileto, o engenheiro austríaco e pai da Terminologia científica, Eugênio (Eugen) Würster, produziu um esquema conceitual sobre o sistemas de conceitos que muito se assemelha ao modelo idealizado pelo filósofo COSTA (2002), sem dúvida influenciado pelo seu pensador predileto, o filósofo Luís Wittgenstein.

A Figura 3.4 ilustra a visão de FELBER (1984), *bipartida* entre o *mundo conceitual* ou do sistema lingüístico e o *mundo real*, das entidades empíricas e da experiência cognitiva humana, dos sons, enfim da *fala*.

A Figura 3.5 ilustra a visão de COSTA (2002), *tripartida* entre o *mundo conceitual* ou do sistema lingüístico, o *mundo real*, da fala, e por uma relação que liga esses dois mundos, a que Wittgenstein denominou de *relação afigurante* (denotação). Esta relação vincula cada nome, que faz parte de uma frase elementar (*Frase elem*$_1$ ... *Frase elem*$_n$), a cada entidade (Ent$_1$ ... Ent$_n$) que faz parte de um estado de coisas ou de um fato elementar (atômico). Esta vinculação é uma perfeita correspondência biunívoca.

Encerrando o processo pictorial da frase narrado por COSTA (2002), os estados de coisas ou fatos atômicos combinar-se-iam entre si para formar os fatos de vários graus de complexidade do mundo. Percebe-se que a relação *afigurante* também delimita dois proces-

sos sucessivos de produção de conhecimento (*modus sciendi*[141]) - a síntese -, que reconstrói o que decompôs o outro *modus sciendi* - a análise.

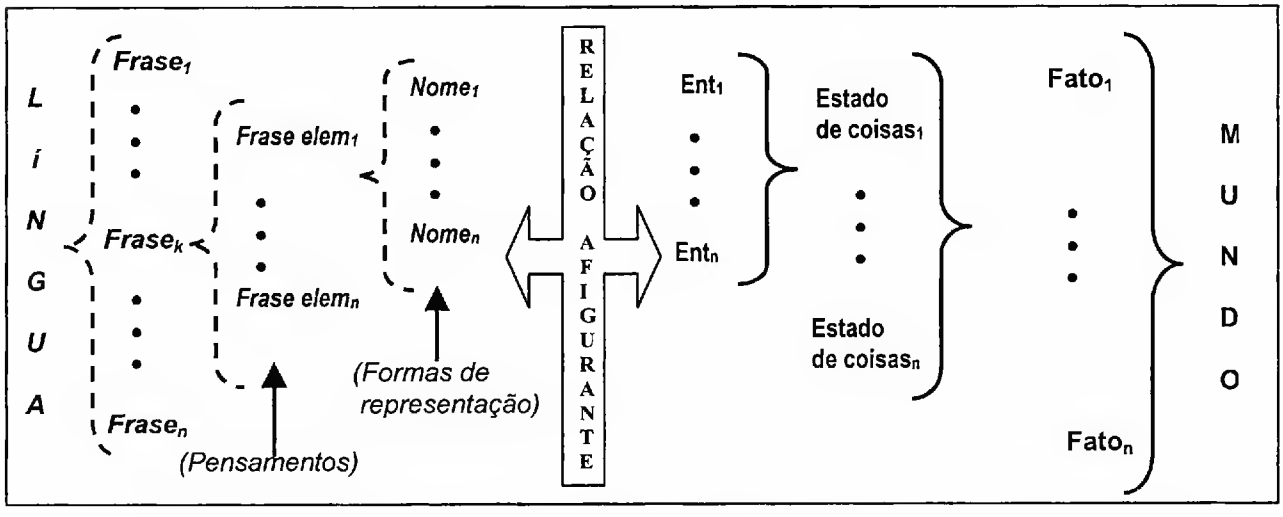

Figura 3.5:Teoria Pictorial da Frase de Wittgenstein segundo COSTA (2002).

Apesar de Wittgenstein ter sofrido muitas críticas, quando passou para a segunda fase de sua produção intelectual, por refletir sobre os problemas da Filosofia da Linguagem Ordinária e por criar controvérsias com relação à sua própria produção intelectual de caráter formal e quase impecável da primeira fase, sua contribuição para as ciências exatas e sociais é incontestável, sendo interessante transcrever as suas impressões pessoais sobre a linguagem, que, de certa forma, coincidem com os cânones de heurística[142] que cientistas da Ciência da Computação [(RUSSELL, 1995) e outros] ou da Lingüística Computacional [(MEDEIROS, 1999) e outros] estabeleceram, a fim de tornar possível o desenvolvimento de sistemas especialistas de PLN, diminuindo os riscos de produzir intratáveis problemas NP. Eis a transcrição[143]:

"A linguagem é como uma nebulosa constituída de múltiplos locais, regiões, sublinguagens mais ou menos aparentadas entre si, e é nelas e nas transgressões de suas fronteiras internas que o filósofo deve focalizar a sua atenção. Mesmo que exista uma unidade geral da linguagem, ela não chega a ser relevante para a investigação filosófica."

[141] Segundo GARCIA (1976, p. 281 e 301), são quatro os métodos subsidiários do raciocínio humano (os fundamentais são a dedução e a indução): análise, síntese, classificação e definição. Os quatro são chamados de *modus sciendi*.
[142] Regra simples que é aplicada durante a solução de um problema, de modo a limitar drasticamente o espaço de busca.
[143] Transcrição que veio por intermédio de COSTA (2002, p.35).

O que se nota de comum nesses esquemas que tentam modelar sistemas complexos do MR em fórmulas tratáveis pela racionalidade são, como dizia POPPER (1975, p. 151), variações do tema do dualismo filosófico ocidental corpo–mente.

As reflexões sobre um terceiro mundo (das idéias) remontam aos platônicos e estóicos gregos, que continuaram em algumas correntes mais modernas, como no racionalismo de Leibniz e no positivismo lógico de Frege e de Wittgenstein, p.ex.

Há muita similaridade entre vários esquemas da Ciência da Computação, da CIGeo e da Lingüística, ilustrados neste trabalho, com a visão dos *três mundos popperianos*: **matéria** (primeiro mundo), **mente** (segundo mundo) e **idéias** (terceiro mundo), em que o mundo da mente provê uma mediação entre os outros dois. É sobre esse tipo de relação mediadora que recai o maior esforço de reflexão dos filósofos pluralistas como Karl R. Popper.

Segundo POPPER (1975, p. 154), os objetos aos quais o ser humano se refere pela linguagem pertencem aos três mundos: no primeiro mundo, pode-se falar sobre coisas ou relações materiais; no segundo, sobre estados mentais subjetivos, juízos de valor sobre uma teoria, i.e., sobre a apreensão dessa teoria; no terceiro, sobre os juízos de fato sobre essa teoria, sobre a interpretação do seu conteúdo, ao avaliar a sua validade.

O que importa sobre esta referência ao trabalho de POPPER (1975, p. 170), neste trecho da revisão, foram as diversas conclusões que tirou, ao refletir sobre as relações do seu universo tripartido e ao exercitar a análise de argumentos no campo das ciências naturais e sociais. Uma dessas conclusões subscreve este subitem e reforça ainda mais as sondagens que se fizeram neste trabalho em áreas alheias à Cartografia:

"Qualquer tentativa (exceto as mais triviais) de compreender uma teoria sujeita-se a abrir uma investigação histórica a respeito dessa teoria e de seu problema, que assim se torna parte do objeto da investigação."

Na visão de PINTO (1977), que é um operador da Semântica, as preocupações com a *significação* e com o modo como as expressões lingüísticas significam e se relacionam com o mundo empírico dos sentidos, ou com o mundo subjetivo dos pensamentos humanos, é tão antiga quanto a filosofia ocidental e remonta aos pré-socráticos.

O enfoque e a análise de PINTO (1977) sobre as três correntes que identificou para Filosofia da Linguagem, apesar de guardarem alguma afinidade com as três principais de COSTA (2002), é fundamental para coroar com argumentação lógica a coerente posição da metodologia adotada por essa linha de pesquisa da CIGeo que estuda a SS.

Para PINTO (1977), as três correntes de Filosofia da Linguagem são as do realismo:

- *Diádico*

- *Triádico*
- Crítico

O triádico, por seu turno, subdivide-se em:

- *Psicologismo*
- *Logicismo*

O psicologismo ainda se subdivide em três visões:

- Visão de C. K. Ogdens e de I. A. Richards (fundo neopositivista)
- Visão de L. Bloomfield (fundo *behaviorista*)
- Visão de L. Witt (fundo pragmático)

O traço marcante do *realismo* ***diádico*** está na *associação direta* entre o significado com a entidade referida no mundo-real (MR) empírico.

No caso do ***triádico***, o significado é identificado como o nexo concebido lógica ou psicologicamente, que assume o papel de mediador entre a expressão e a entidade referida.

No caso do ***crítico***, o significado é uma propriedade intrínseca à expressão lingüística[144], inseparável dela. É o significado que determina a possibilidade de se usar a linguagem com funções de aspecto simbólico, como referenciar entidades do MR, expressar pensamentos ou provocar comportamentos.

PINTO (1977) admitiu que parte da produção intelectual dos três filósofos[145] apreciados por COSTA (2002), anteriormente, esteve inserida no realismo crítico.

Ao analisar o realismo diádico (primitivo ou ingênuo), PINTO (1977) o caracterizou pela confusão que cria entre significado e entidade referida pela expressão lingüística.

Tanto Frege como Russell foram atraídos por essa corrente numa etapa de suas reflexões, particularmente pela simplicidade que lhe é inerente. O problema é que a coisa referida não deveria pertencer apenas às classes de coisas concretas, mas também a outras mais abstratas, como emoções, seres imaginários, etc., o que não chegou realmente a abalar esses dois filósofos, acrescentou PINTO (1977). Essa limitação, causada pela postura materialista do realismo diádico, desorienta uma teoria semântica para as LNs e afasta a maioria dos filósofos e cientistas dessa corrente.

Por essa razão, PINTO (1977) levantou a crítica de insustentabilidade dessa corrente de pensamento filosófico para fundamentar uma teoria semântica para as LNs, visto que os esforços que se fariam para delimitar o que deveria constituir uma entidade empírica derrubaria, no mínimo, a generalização de uma teoria. Exemplificando: só seriam aceitas expres-

---

[144] Classe de ocorrências concretas de sinais (*significante*).
[145] O autor também incluiu Saussure e Charles Peirce.

sões com sintagmas nominais que contivessem pronomes ou artigos definidos e os substantivos que formassem o núcleo desses sintagmas deveriam pertencer à classes de entidades concretas (nada de bruxas, dragões e unicórnios).

Como se vê, são muitas as restrições que descartam o postulado do realismo diádico do campo de problemas de uma teoria semântica. Uma teoria semântica não pode se apoiar na aceitação de que a relação de designação entre a expressão lingüística e a entidade referida tem pertinência para a Semântica, uma vez que o ato de designar não cria significado, nem transforma entidades extralingüísticas em significado (PINTO, 1977).

No triádico, a expressão passa a ser vista como um representante da entidade (papel intermediário ou mediador), ocupando o lugar deste "por delegação" do usuário da linguagem, num processo de simbolização[146] (PINTO, 1977).

O traço comum de todas as visões do psicologismo é a insistência em associar à teoria do significado a explicação de que este é produzido pelo falante ou intérprete, quer dizer, o recurso de produção individual derruba o caráter de interpessoalidade que preside a comunicação. A tradição semântica, segundo PINTO (1977), estabelece que a seguinte fórmula (e não a recíproca) é verdadeira:

Significado ⇒ conhecimento, valores, intenções, desejos

A fórmula esquemática acima quer dizer o seguinte: Só se pode referenciar entidades do MR, expressar pensamentos ou provocar comportamentos, se a expressão lingüística for dotada de significado (como se verá, esta é a síntese dos postulados do realismo crítico).

A visão de Ogdens e Richards foi lamentavelmente muito citada e recomendada em manuais de lingüística e até em livros didáticos do nível secundário, segundo PINTO (1977). O lamentável advém do fato de que a visão se torna vulnerabilíssima, se aplicada fora das línguas formais.

De qualquer forma, se aplicada em línguas profissionais ou na Terminologia, a visão de Ogdens e Richards é capaz de desencadear algumas hipóteses de trabalho interessantes em estudos-de-caso como os de RODRÍGUEZ (2000) e o desta tese, tanto que o triângulo desses pesquisadores guarda alguma semelhança com o de DAHLBERG (1978), de extensa aplicação em trabalhos terminológicos.

É fácil visualizar o esquema de Ogdens e Richards. Na base de um triângulo esquemático, ficam os vértices denominados **símbolo** (expressão) e **referente** (entidade do MR), ligados por uma linha tracejada (relação convencional de equivalência). Ambos os vértices

[146] A designação, segundo PINTO (1977), é um processo de simbolização.

são apontados por dois vetores que se originam do terceiro vértice – o **pensamento** (*engrama*). Esses vetores são relações causais, i.e., equivalem a relações lógicas de implicação. Segundo PINTO (1977), o pensamento resulta de experiências sensíveis acumuladas durante a vida dos indivíduos e que desencadeiam (implicam, acarretam) relações formais, espaciais e temporais entre diferentes entidades e eventos que participam de uma situação.

Para entender a arrazoada crítica de PINTO (1977) ao modelo em pauta, é preciso recordar que uma asserção de implicação **p** ⇒ **q** ("p" implica "q") significa que a proposição antecedente ("p") é uma condição suficiente para o conseqüente ("q") e que este é uma condição necessária decorrente de "p". A implicação só é falsa se "q" for falso diante de um "p" verdadeiro. No campo empírico, segundo RUDIO (1978), uma condição é necessária, quando, sem ela, o fenômeno não pode ser produzido. E uma condição é suficiente, quando, se presente, produz inevitavelmente o fenômeno.

Aplicando esses fundamentos à visão de Ogdens e Richards, PINTO (1977) localizou uma falha básica do modelo, confrontando-o com as evidências semânticas já existentes. Pelo modelo, a existência empírica da entidade do MR (referente) é necessária e decorrente da formação do pensamento positivo. Até aí, o modelo é logicamente plausível. Mas quando a relação de equivalência entre referente e símbolo se estabelece, ou seja, quando só pode haver símbolos se houver referentes por ele simbolizados e vice-versa, aí o modelo cai num *monossemismo* irreal, em que não existe ambigüidade nesse processo de representação.

A visão de L. Bloomfield (escola americana de lingüística), de fundo comportamentalista, rejeitava qualquer recurso a entidades mentais, por não serem observáveis empiricamente. Esse enfoque determina um estreitíssimo horizonte de observação para os fenômenos lingüísticos (semânticos). Sendo assim, esse modelo só admite estudos do significado expressos por enunciados imperativos e performativos (que denotem ações de resultados mensuráveis). PINTO (1977) concluiu que foi por causa dessas avaliações estreitas e reducionistas do modelo *bloomfieldiano*, e dos malogros resultantes, que os lingüitas norte-americanos desistiram dos estudos do significado por mais de três décadas.

Na visão pragmática de L. Witt, o significado é determinado pelo uso que o falante faz das expressões. Este uso é de características muito pessoais e derruba o princípio interpessoalidade da comunicação. Como observou PINTO (1977), o significado seria totalmente dependente do contexto da fala.

O ***logicismo triádico*** foi considerado por PINTO (1977) como uma das linhas de pensamento mais promissoras da Filosofia da Linguagem, para a criação de uma teoria semântica. O principal motivo de sua resistência em relação às outras linhas é que nela se verifica

um esforço permanente para manter o princípio da interpessoalidade da comunicação, preservando um ***conceito coerente de significado***. Para isto, como disse PINTO (1977), existe um entendimento dos *logicistas* para localizar o estudo do significado em dois campos totalmente independentes: o da Semântica e o da Pragmática.

No estudo do significado pelo campo da Semântica, os resultados se mostraram mais promissores, visto que a Lógica e a Matemática norteiam este estudo, que se utiliza de conceitos como o de *denotação*[147] e de *conotação*, que vêm de J. Stwart Mill (1843), segundo PINTO (1977).

No caso da Pragmática, que cuida do uso individual da linguagem, é mais complexo manter a uniformidade que se tem para o conceito de significado em relação ao campo da Semântica.

Na Pragmática, os desdobramentos do estudo avançam os campos de outras ciências cognitivas, particularmente a Psicologia Cognitiva. A IA tem sido a disciplina que mais tem operado nessa zona multidisciplinar para formar evidências e conhecimento que coadunem os objetos de todas as ciências envolvidas. Trabalhos como os de MILLER (1991), no campo da Lingüística, e de MEDEIROS (1999) e RODRÍGUEZ (2000), no campo da IA, ao focalizarem o contexto em que ocorre a comunicação, são provas desses esforços.

Para encerrar este subitem, serão apreciadas algumas considerações críticas de PINTO (1977) sobre a corrente do *realismo crítico*, partilhada por Wittgenstein, Saussure e Charles Peirce, em várias ocasiões de suas produções intelectuais.

Essa corrente emana de três postulados:

- P-1: "O significado é intrínseco à expressão lingüística".
- P-2: "A Filosofia da Linguagem é uma filosofia do senso comum, porque só por intermédio da linguagem o homem tem acesso ao conhecimento".
- P-3: "O conhecimento é um repositório estruturado de construções lingüísticas e de definições (delimitações de conceitos) sobre objetos de qualquer natureza (concretos ou imaginários)".

Para Saussure e Wittgenstein ficou bem claro que o significado é imanente à expressão lingüística, independente do mundo empírico, e possui um valor dentro do sistema de LN à qual pertence a expressão (termo). Donde se conclui que, se esse valor está inserido num sistema, é possível medi-lo, ou mais precisamente: avaliá-lo relativamente a outros va-

---

[147] Também denominada de *referência hipotética* ou *aplicação*. *Denotação* e *conotação* serão descritas a seguir.

lores (nível da *conotação*); avaliação de valor que a tese de RODRÍGUEZ (2000) e a presente tese estabeleceram alcançar como objetivo.

Para Saussure e Wittgenstein ficou a convicção de que também existe uma relação chamada de *significação*, que associa os pensamentos individuais dos falantes à expressão lingüística.

A tarefa de avaliar o valor relativo da significação é mais complexa, porque ela já pertence ao campo de problemas psicológicos. Mas nem por isso cessaram as tentativas de mensurar o seu valor: o trabalho de RODRÍGEZ (2000) também lidou com ontologias que levaram em consideração o contexto e o discurso. São trabalhos dessa natureza que põem à prova o já mencionado ***princípio leibniziano da intersubstituibilidade***, o qual se manifesta no nível da *denotação* ou dos vários pensamentos ou significações individuais.

Por conseguinte, conhecer ("medir") o significado de uma expressão pelo método denotativo é o mesmo que conhecer a diferença de emprego entre esta expressão e outras do mesmo sistema de LN.

No nível da avaliação do *valor do significado* pelo *método conotativo* e do *valor da significação* pelo *método denotativo*, é mais plausível a fixação do valor do significado do que do da significação, no estágio atual das pesquisas.

Esses cinco últimos parágrafos já fazem um recorte importante para o referencial teórico desta tese.

Apontamentos sobre as teorias do conceito: o subitem anterior serviu para rever muitos conceitos ligados ao campo da Lingüística e, particularmente, os ligados à Semântica, o que será de muita valia para tratar de considerações terminológicas mínimas para a montagem das definições que serão utilizadas na formulação de ontologias para a carga no PRONTO©.

O que vem a seguir, até o término do subitem, trata de conceitos muito complexos e muitas vezes confundidos uns com os outros, tais como: **signo** *(significante e significado)*, **significação**, **conceito** (suas relações), **denotação**, **conotação** e outros mais palpáveis como **definição**. Alguns deles já foram antevistos de forma superficial.

Antes da total aquisição da faculdade da linguagem, o homem é incapaz de perceber racionalmente quaisquer entidades do mundo que o cerca. O simples contacto sensorial com as entidades empíricas não é suficiente para que ele detecte os significados (sentidos) que as coisas e suas relações entre si revelam e, por conseguinte, para que produza conhecimento. E é esta faculdade da linguagem, adquirida ao longo de penosos e fortuitos proces-

sos evolutivos, que o habilita a categorizar[148] as coisa e relações entre as coisas do mundo, i.e., de criar **sistemas de conceitos**.

O que se conhece de uma entidade percebida no MR são apenas as suas características distintivas (traços ou feições distintivas), que a relacionam com outras que com ela começam a formar um sistema conceitual no aparelho cognitivo humano.

É por isso que Frege diferenciou *característica* de *propriedade* (V. glossário). Para ele, o essencial poder de generalização repousa nas características das coisas ou entidades. Essas características permitem organizar as entidades do MR por classes ou categorias. Esse poder torna as características mais limitadas quantitativamente, ou melhor, "mais enumeráveis" do que as propriedades dessas coisas, que não desfrutam do poder de categorização, mas sim do de descrição das instâncias das classes de coisas. Se não forem especificados limiares para esta enumeração, a quantidade de propriedades pode tender ao infinito.

Assim como a linguagem desencadeia (acarreta) os sistemas de conceitos, o significado, imanente à linguagem, produz (acarreta), como já visto, conhecimento, valores, intenções e desejos.

Mas antes de tratar mais especificamente do sistema de conceitos, é preciso investigar um conceito primordial ("o conceito dos conceitos"): o **signo**.

Para esse conceito, foi necessário consultar críticos de Lingüística (e Semântica) como DUCROT (1972) e PINTO (1977), bem como autores de Filosofia (MORA, 1994), sem mergulhar no âmago de alguns pontos muito controvertidos. O intuito foi o de, tão-somente, extrair desses autores o que for de consensual e tangível para um pesquisador de fora desses campos, para fundamentar o seu problema.

Pelas descrições, reflexões e algumas definições vistas até este ponto, já é possível aproximar-se do conceito de **signo lingüístico**.

Segundo Saussure, os **signos** lingüísticos são **tangíveis**; a escrita pode fixá-los em imagens convencionais. Ao se comparar o signo lingüístico com os pormenores do ato da fala (mesmo concretos), esbarra-se na impossibilidade de materializar esses pormenores em algum suporte, p.ex., na impossibilidade de "congelá-los numa fotografia".

A língua e a fala são objetos de natureza concreta, diferentemente da linguagem, no entanto, na língua não existe senão a ("imagem") acústica, que pode traduzir-se numa reprodução visual constante. É essa possibilidade de fixar as coisas relativas à língua que faz

---

[148] Segundo PINTO (1977), a faculdade de classificar é o "sujeito transcendental de Kant".

com que um dicionário e uma gramática possam representá-la fielmente. A língua é, portanto, um depósito das "imagens" acústicas; a escrita, a forma tangível dessas imagens.

Se fosse possível abarcar a totalidade das imagens verbais armazenadas em todos os indivíduos, seria possível atingir o limite do que seria realmente a língua, já que a língua não está completa em nenhum dos cérebros dessa massa humana (SAUSSURE, 1975).

Segundo DUCROT (1972), o signo é a noção básica de toda a ciência da linguagem; porém, em virtude dessa própria importância, é uma das mais complexas de definir. Essa complexidade aumenta pelo fato de que se tenta, nas modernas teorias do signo, considerar não apenas entidades lingüísticas, mas também signos não-verbais. O autor afirmou que as definições clássicas do signo, em acurado exame, mostram-se, amiúde, ou tautológicas (recursivas), ou incapazes de apreender o conceito em sua essência.

Ao refletir sobre a natureza do signo (lingüístico), SAUSSURE (1975) induziu uma importante propriedade dessa entidade abstrata: "O **signo lingüístico** une não uma coisa e uma palavra, mas um conceito e uma imagem acústica[149]". Esta natureza dual e inseparável desses dois elementos (conceito/**significado** e imagem acústica/***significante***) é que introduz a mais expressiva questão lingüística em torno do termo **signo**, elemento amalgamador do *significante* e do significado, para o qual o pensador genebrino enunciou dois princípios: o da *arbitrariedade do signo* e o do *caráter linear do significante.*

Pelo primeiro princípio, então, a idéia de "árvore" não está ligada por relação alguma interior à seqüência de sons "á-r-v-o-r-e", que lhe serve de *significante*; poderia até ser representada igualmente por outra seqüência, como provam as diferenças lingüísticas existentes para a idéia de árvore ("t-r-e-e" no inglês, "a-r-b-r-e" no francês, p.ex.).

No caso do segundo princípio do signo lingüístico, Saussure ressaltou a natureza auditiva do *significante*, que se desenvolve no tempo (gradativamente, como numa progressão linear) e que pode acumular novas características.

Dentro do princípio da linearidade do *significante* do signo lingüístico, Saussure fez menção aos ***significantes* visuais**, dando especial ênfase à complexidade deste caso, em virtude da *multidimensionalidade*[150] que está presente nos **signos gráficos**, os quais não acumulam novas características conceituais apenas na linha única do tempo, mas também nas dimensões espaciais.

---

[149] Esta imagem acústica é a representação natural da palavra como fato da língua, que transcende (mas inclui) os domínios da fala (esforços musculares para produzir a fonação, etc.).

[150] V. subitem 1.5.1, conforme as visões de BERTIN (1967) e MARTINELLI (1991).

É interessante notar como esses autores de ensaios, artigos e livros de Lingüística, Semiótica e Semântica, na sua essência, já delineavam, há muitos anos e até décadas antes do surgimento dos métodos de modelagem orientada a objetos[151], os requisitos para criar linguagens técnicas de programação (LTPs) que pudessem capturar as relações entre entidades espaciais com mais realismo, em què pese os primeiros autores serem raramente (ou nem serem) citados pelos criadores dessas técnicas, talvez até pelo que concluiu SÖRGEL (1999)[152] sobre a falta de sinergia entre essas áreas.

PINTO (1977) ressaltou que o *significante* de Saussure não deve ser associado à ocorrência de um objeto empírico, mas sim a um tipo (*classe*) de objeto empírico, i.e., por um conjunto arbitrário de sucessões de sons ou de gráficos (mapas) que teriam o mesmo uso para o falante ou usuário de um sistema de linguagem, quer dizer, essas sucessões de sons ou gráficos serviriam para veicular (transportar) o mesmo significado, que, por sua vez, seria um tipo (classe) arbitrário de pensamentos que poderiam ser veiculados (transportados) pelo mesmo *significante – o* portador de conteúdo informativo (o significado).

Essas explanações são importantes para delimitar as noções equivocadas que se têm de signo, normalmente associando-o a uma determinada manifestação acústica ou à uma certa representação gráfica (que é uma instância ou ocorrência singular do *significante*).

PINTO (1977), por conseguinte, estabeleceu um divisor d'água importante para o *sinal concreto* (sons e gráficos), de natureza variável, e o *significante*, conjunto arbitrário de sinais invariáveis da linguagem (sistema lingüístico).

Do mesmo modo, distinguiu um ***pensamento isolado*** (individual) associado ao sinal concreto, a que chama de ***significação*** (campo da fala e da representação gráfica que um indivíduo executa num sistema de linguagem), do ***significado***, ***conjunto arbitrário de significações*** (sistema de linguagem).

Assim, o signo estaria numa posição transcendente em relação ao mero sinal. O *signo* seria a *sublimação do sinal concreto* e este é a marca perceptível do signo: ***"O sinal representa[153] o significante e apresenta o significado"*** (PINTO, 1977). De forma mais abrangente: A língua apresenta o mundo, quando suas expressões (frases ou termos isolados) são portadoras de informação significativa e, por outro lado, representa esse mundo, quando utiliza funções *sígnicas* (*iconização*, *indiciação* e *simbolização*) para exprimi-lo na comunicação.

---

[151] MMOOs (V. glossário): OOSE™ de Ivar Jacobson, OOD™ de Grady Booch e OMT™ de James Rumbaugh, das quais, por intermédio dos esforços do OMG, surgiu a UML™ (SILVA, 2001).

[152] V. subitem 1.5.3.1.

[153] **Apresentar**: pôr na presença; **representar**: reproduzir por meio de imagem (LELLO, 1984).

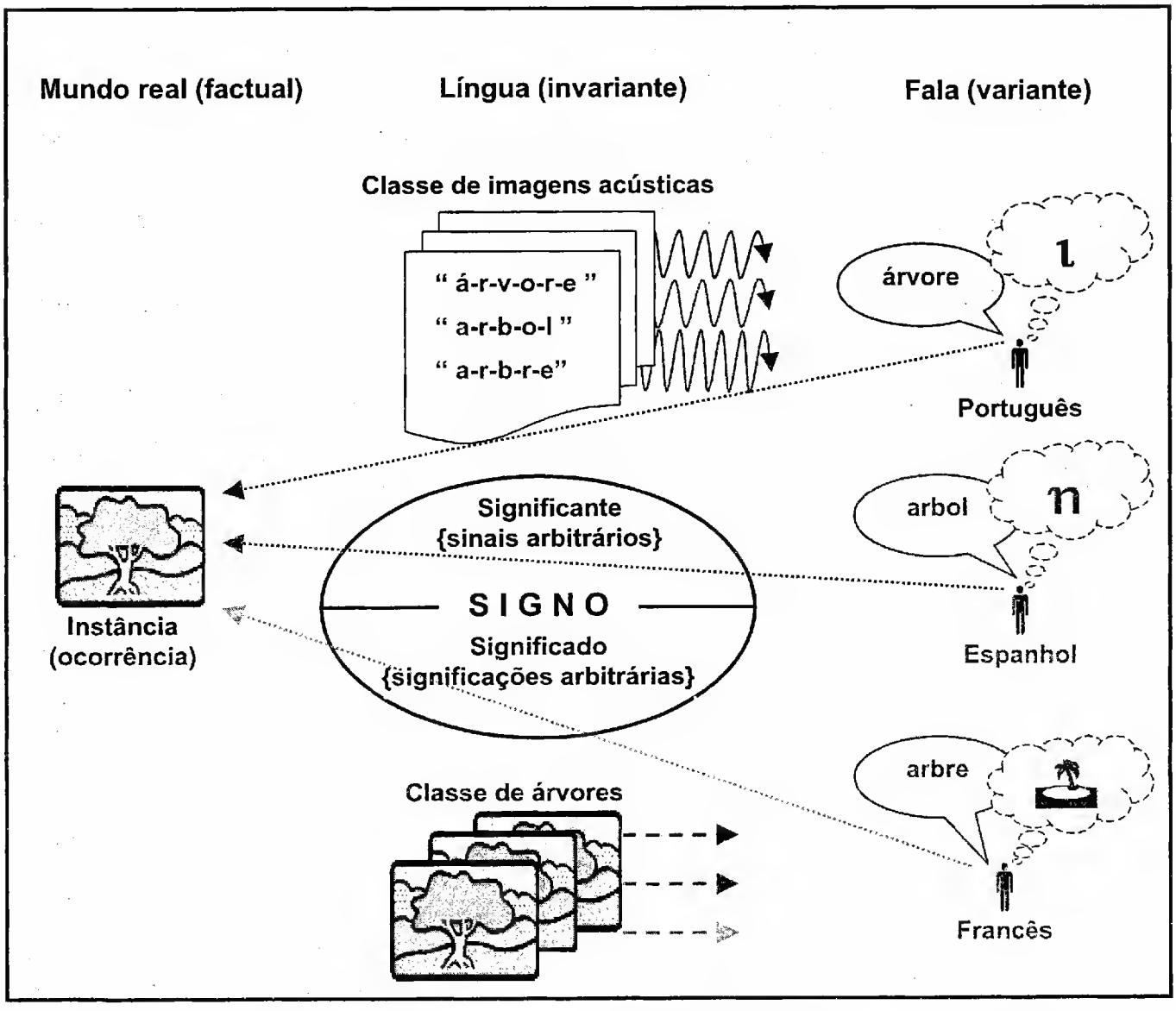

Figura 3.6: Apresentação e representação da realidade por um sistema de linguagem.

PINTO (1977) ainda desejou eliminar lacunas de entendimento, apoiando-se em importantes pesquisadores e filósofos[154], quanto ao que poderia ter pensado Saussure, quando disse que o **signo** é **arbitrário** ou *imotivado* e que o **símbolo** (sinal concreto) é **convencional**, motivado ou não-arbitrário e que a constituição das classes abstratas ou formais chamadas *significante* e significado não depende de nenhum motivo (imotivado) intrínseco à matéria sonora (gráfica) ou psicológica que agregam. De outra forma: a formação do sistema da língua está sujeita a regras internas que lhe são próprias e não à pressão da realidade empírica, externa ao sistema.

[154] T. de Mauro, É. Bienveniste, Umberto Eco e R. Jakobson.

A Figura 3.6 esquematiza o conceito de signo no campo da linguagem e no da fala, explicitando os processos de *apresentação do significado* e de *representação* (simbolização, *indiciação* ou *iconização*) *do significante* num sistema de linguagem.

O esquema está dividido em três "mundos"[155] ou domínios: o real, o da língua (estável) e o da fala (variável). Cada cor representa um tipo de conjunto arbitrário de sons – fala: azul para o português, vermelho para o espanhol e verde para o francês – e um conjunto arbitrário de pensamentos individuais (significações) de cada um dos três falantes.

Na Figura 3.6, uma entidade do mundo real (uma ocorrência de árvore) está sendo observada por três falantes de cada idioma, que pronunciam a seqüência arbitrária de sons que denotam o conceito de árvore (classe dos objetos com as características de árvore).

No domínio da fala, vêm ilustrados os sinais concretos (que se ouvem) e variáveis de cada sistema fonético.

Os significantes "á-r-v-o-r-e", "a-r-b-o-l" e "a-r-b-r-e" não devem ser confundidos com a ocorrência de uma certa árvore no mundo real. São conjuntos arbitrários de sucessões de sons ou de gráficos que têm o mesmo uso para os falantes de cada idioma e que conotam as características essenciais da classe de objetos conhecida como *árvore*.

As setas coloridas e tracejadas apontam para os processos de representação mental (pensamentos, significações) de cada falante. Esses processos vêm ilustrados como névoas que se originam de cada subdomínio fonético (falante de um idioma).

Ao contrário da fala, o domínio da língua é um *invariante*, que se situa num plano abstrato (onde se cruzam todos os raios de observação coloridos da Figura 3.6), chamado de *signo lingüístico*, formado por entidades mentais inseparáveis: o *significante* e o significado.

As linhas sinuosas coloridas traduzem os processos de simbolização, indiciação ou de iconização, que associam o **sinal** - as marcas perceptíveis do signo em cada sistema fonético - ao *significante* (o sinal gráfico "árvore" representa o *significante* e, nessa fase, o mesmo sinal gráfico apresenta o significado, conforme já explanado).

Para PINTO (1977), Frege foi o lógico-matemático que mais aproximou a sua metodologia das necessidades de uma teoria semântica para LNs, em que pese todas as limitações da sua Teoria do Sentido. Realmente, o exemplo do planeta Vênus, referido por "estrela da manhã" e por "estrela da tarde", é um primor de explanação para frases de identidade e do

---

[155] Mais uma vez, a tripartição de mundos, lembrando POPPER (1975).

seu poder de portar ou não informação significativa. O esquema para esta parte da teoria de Frege seria o seguinte[156]:

$$sinal_1 \prec sentido_1 \prec referência \succ sentido_n \succ sinal_n$$

O sinal não está no lugar do objeto real (referência); ele representa uma classe genérica desse objeto por meio de uma sucessão sistematizada de fonemas de um sistema lingüístico e apresenta essa classe genérica de objetos por meio (mediação) do significado (sentido de Frege), que é imanente ao sinal.

Há ainda alguns conceitos muito bem firmados na Filosofia e na Lógica, que acabaram passando para a IA e, dentre eles, alguns merecem ser definidos, como: ***esquema***, ***homogeneidade*** e semelhança ou ***similaridade***, todos à luz de MORA (1994).

O sentido de **esquema**, p.ex., remonta aos clássicos gregos, normalmente ligado às noções de **forma** e **idéia**.

Para Kant, o conceito de *esquema* era o de uma representação homogênea e mediadora[157] entre a categoria (idéia) e a aparência (forma) do fenômeno, tal que fosse possível a aplicação da primeira (categoria) à segunda (aparência).

Apesar da confusão que habitualmente se faz entre *esquema* e *imagem ou representação mental*, há pequenas diferenças entre esses conceitos na Filosofia. Em primeiro lugar, um esquema é que determina uma imagem. O primeiro é um produto da imaginação pura; é um procedimento universal da imaginação humana. O segundo, não. A imagem é um produto da imaginação empírica, provocada por estímulos externos à imaginação.

O conceito de *esquema* foi aproveitado nas mais modernas teorias do conceito, como se verá no subitem seguinte, particularmente pelo fato de propiciar a sua *quase-formalização* por intermédio de ontologias.

A homogeneidade é um conceito que esteve originalmente ligado a duas coisas que pertencessem ao mesmo gênero e que compusessem uma outra coisa mais complexa (o composto).

Passando por Kant, a homogeneidade passou a ser um atributo ligado à aparência das coisas submetidas a um processo de ordenação hierárquica. Segundo a visão kantiana, a representação de um objeto (um prato, p.ex.) deve ser homogênea em relação ao conceito que a ele se aplica (uma abstração matemática, como um círculo, p.ex.).

---

[156] Os símbolos $\prec$ $\succ$ significam sucessão e precessão, respectivamente, numa relação de ordem.
[157] Mais uma vez uma semelhança com os três mundos *popperianos*.

As implicações da homogeneidade kantiana nas teorias do conceito são muito amplas, sempre ligadas à possibilidade de se fazer síntese, *modus sciendi* que é subsidiário da razão para disciplinar o aparelho intelectual humano na aquisição de conhecimento.

A síntese, para Kant, é a unificação de algo que seja originariamente diverso (heterogêneo), no momento da observação[158]. O ser humano, então, exerce suas faculdades cognitivas sobre a diversidade, para começar a construir um sistema de conceitos. Durante este esforço de construção, são capturadas homogeneidades na diversidade original, as quais são instanciadas em conceitos de coisas empíricas, sobre os quais se pode exercitar a definição (limitar o conceito), experimentar e descobrir outras diversidades e, daí, recomeçar todo o processo, com a criação de novos conceitos, num processo interminável. O conceito seria, por conseguinte, a matéria-prima para a razão humana.

Para montar uma estrutura de RC do grupo semântico fraco ou forte (V. Figura 1.13), é preciso entender este ciclo de produção de conceitos, que abrange três etapas: 1ª) **Extrair** categorias fundamentais e gerais (homogeneidade) da diversidade original; 2ª) **Derivar** categorias mais específicas da variedade de homogeneidades obtidas; e 3ª) **Montar** novas unidades das homogeneidades por meios que reproduzam a lei de afinidade de conceitos.

O princípio que rege o esforço de construção de um sistema de conceitos é o da ***economia cognitiva***, i.e., a razão divide o MR em classes de coisas para diminuir a quantidade de informação de que necessita para processar (é a análise, outro *modus sciendi*). Este processo foi exaustivamente descrito por QUILLIAN (1968).

Outro conceito que será mais esmiuçado no subitem 3.2.2.2.2, mas de forma quantitativa, é o da similaridade.

Neste ponto, apenas a noção filosófica dada por MORA (1994) será descrita.

Para a Filosofia, semelhança e *similaridade* são sinônimos, pertencendo a um grupo de outras relações homogêneas como a *igualdade*, a *identidade* e a *diferença*.

Grosso modo, pode-se dizer que duas entidades são similares entre si, quando não são idênticas, nem iguais e nem diferentes (embora possam ser homogêneas), mas possuem, simultaneamente, algo em comum e algo distinto entre si, ou seja, duas entidades podem ser similares sem pertencer à mesma espécie.

A noção de similaridade ocorre em três contextos, na Filosofia:

- Dada uma entidade "A", a similar "a" é uma especificação da primeira;
- Dada uma entidade "A", a similar "A' " é uma cópia exata da primeira por referência a ela;

[158] Segundo RUDIO (1980), são três os níveis de intervenção do homem na realidade, durante uma pesquisa: a experiência (nível da simples atenção), a observação (atenção mais rigorosa) e experimento (interferência direta no fenômeno).

- Dada uma entidade "A", a similar "B", apesar de não ser da mesma espécie de "A", possui elementos estruturais semelhantes aos da primeira.

A OO também trata do primeiro contexto (relação de generalização) e do terceiro contexto (relação de agregação), por um prisma de aplicação da teoria à prática.

Fugindo momentaneamente da órbita da Filosofia pura e voltando para um contexto um pouco mais pragmático do assunto, serão revisadas as características das principais teorias do conceito, mas com intenção de emprego prático.

Problemas de terminologia, essenciais para a construção de glossários e trabalhos terminográficos, deixaram de ser exclusivos da Lingüística e da Semântica e passaram a compor os quadros referenciais das *geociências*.

São novas comunidades de usuários que necessitam definir novos códigos de comunicação para tratar da complexa IG. A trilogia de CÂMARA (2002) é um indicador dessa tendência. E é sobre essa tendência que devem acorrer lingüistas e estudiosos de semântica, para contribuir com seus colegas das ciências naturais na solução desses problemas genuinamente interdisciplinares.

Um desdobramento dessa convergência de objetos disciplinares é o surgimento de campos de pesquisa como a Lingüística Computacional e a CIGeo, que obviamente despertarão novos temas de reflexão para a Epistemologia e para a Filosofia da Linguagem.

Um meio-termo entre todos os autores modernos que se preocupam com o "conceito de conceito", grande parte deles oriunda das ciências sociais, é o fato de se despirem do vício herdado do positivismo, em que se desconsidera a explicação científica funcional ou *teleológica* das ciências sociais e do homem em geral (COUTO, 1983).

Aplicar modelos matemáticos das ciências formais aos problemas sociais, sem a necessária adequação racional aos limites do problema, é degenerar num positivismo vazio, almejando-se, em vão, que trabalhos desse tipo adquiram um grau maior de confiabilidade.

E para a aplicação mais judiciosa desses modelos, é primordial conhecer alguns conceitos e definições direta e indiretamente ligados ao fenômeno de interesse neste trabalho: a similaridade semântica (SS), situando-a no quadro referencial ou no marco teórico de uma das várias teorias do conceito.

Segundo HENRIQUES (2001), há duas grandes correntes para as teorias do conceito:

- A que se orienta por modelos de organização simples do conceito;
- A que se orienta por modelos de organização complexa do conceito.

A primeira tem o objetivo geral de explicar como as diversas entidades do MR podem ser grupadas num conceito comum e como se relacionam entre si. É a linha predileta de au-

tores que desejam formalizar os seus trabalhos ao empregar modelos matemáticos e ao evitar, tanto quanto possível, que suas teses tratem de aspectos qualitativos mais complexos.

Quadro 3.2: Teorias do Conceito (Organização Simples)

Atributos de Definição	Protótipo
- Oriunda da Filosofia e da Lógica	**- Oriunda de pesquisas da ψ Cognitiva**
conceito ::= {atributos} (Gottlob Frege)	*princípio da tipicidade ::= f (tempo resposta)*[159]
- *Intensão* e *extensão* do conceito	**- *As* classes de coisas do MR são organizadas com base em protótipos centrais (típicos) ou pelos melhores exemplos do conceito.**
- O significado de um conceito pode ser capturado por uma lista de atributos conjuntivos (≈ *consins*[160]) ou unds. básicas do conceito.	**- O protótipo central é representado pelos atributos característicos que possuem diferentes pesos e graus de relação dentro do conceito.**
- É suficiente para que um conceito seja definido que exista um conjunto de atributos.	**- Uma entidade é membro de um conceito, se seus atributos são similares aos do protótipo (dentro de um critério determinado).**
- Todos os membros de um conceito são igualmente representativos.	**- Uma entidade ∈ uma classe, se ela é similar aos melhores exemplos (ocorrências, instâncias) do conceito.**
- Quando são organizados numa taxinomia, os atributos de um conceito específico vêm do mais geral (herança).	**Um conjunto de atributos é necessário mas não é suficiente para definir um conceito.**
- Os conceitos classificam as coisas do MR e os limites entre as classes dessas coisas são muito rígidos e bem definidos.	**- Determinar os atributos de uma entidade é o 1º passo para mensurar a similaridade desta entidade com o protótipo de uma classe.**
- Há mais atributos comuns entre um conceito e o seu ancestral imediato do que entre este conceito e um ancestral mais afastado.	

[159] O gradiente de tipicidade foi medido por L. J. Rips e por E. Rosch, com base nos tempos de reação de indivíduos expostos a tarefas de categorização.

[160] Na metodologia (Cap. 6), *consin* é a abreviatura para uma lista ou conjunto de sinônimos de termos.

A segunda tem o objetivo geral de explicar como os sistemas conceituais (conceitos isolados ou em grupo) se estruturam e são utilizados pelas diversas capacidades cognitivas do ser humano (percepção, atenção, memorização e linguagem, p.ex.).

Para ilustrar as principais características das duas teorias da primeira corrente, que mais contribuíram com o MSS de RODRÍGUEZ (2000) e, por extensão, com o PRONTO©, o Quadro 3.2 é indicado para apreciação. Essas teorias são:

- Teoria dos Atributos de Definição;
- Teoria do Protótipo

O Quadro 3.3 complementa o Quadro 3.2, citando as críticas a ambas as teorias (as características em vermelho do Quadro 3.2 são as sujeitas a críticas). É bom frisar que, cronologicamente, a Teoria dos Atributos de Definição é a mais antiga (vem de Aristóteles, passando dor Kant!) e que a Teoria do Protótipo de Eleanor Rosch (década de 60) tentou suplantar algumas lacunas da anterior, por intermédio do ***princípio da tipicidade***, mas nem sempre com sucesso pleno.

Não é necessário deter-se demasiadamente sobre a segunda corrente. Das novas teorias que vêm surgindo dessa corrente, vale a pena citar a **Teoria do Roteiro** de R. Schank [*apud* HENRIQUES(2001)]. Essa teoria vêm recebendo especial atenção no meio científico, mercê dos diversos pesquisadores que sobre ela se debruçaram, como L. J. Rips, D. E. Rumelhart, D. Gentner, A. Ortony, Linda Coleman, Paul Kay e muitos outros. Eles vêm realizando testes de validação para os quais os resultados têm demonstrado coerência e consistência da teoria com as evidências empíricas.

Quadro 3.3: Críticas às Teorias do Conceito (Organização Simples)

Atributos de Definição	Protótipo
- Foi provado que nem todos os atributos são igualmente relevantes, salientes ou proeminentes.	**- Nem todos os conceitos têm características de protótipo (conceitos abstratos, p.ex: crença, angústia, etc.).**
- Nem sempre é suficiente a definição de um conjunto de atributos para que um conceito seja definido.	**- Dá mais valor aos atributos do que às relações entre as classes às quais eles pertencem. Essas relações são do domínio cognitivo das pessoas.**

Em que pese as teorias da segunda corrente ainda estarem sob testes de validação, tem havido um fluxo de novos conceitos (*constructos*), como o de ***esquema***, p.ex., para as mais simples e tradicionais teorias da primeira corrente.

A definição de ontologia parece trazer muita carga teórica da definição de esquema, que possui a característica de ser maleável o suficiente para acomodar a representação de diferentes tipos de conhecimento. Outras características dos esquemas são:

- Abrigar diversos tipos de relações, indo além das tradicionais relações de gênero-espécie e de parte-todo;
- Codificar conhecimento genérico, que pode ser aplicado a distintas situações específicas;
- Subdividir-se em ***subesquemas*** por intermédio do conceito de ***variável*** ou de ***fenda***, que organizam o conhecimento em situações estereotipadas (p.ex: um homem poderia assumir ***papéis*** como o de cozinheiro, empregado e diretor, todos acumulados numa variável ou fenda específica).

As pequenas resistências à Teoria do Roteiro de Schank obrigaram o autor a fazer complementações, que a enrobustecem cada vez mais. Essas incorporações são *constructos* chamados de *pacotes de organização da memória* (POM ou MOP, em inglês) e de *pontos de organização temática* (POT).

Em poucas palavras, os POMs e POTs têm a finalidade de cobrir parte das lacunas que algumas teorias baseadas na Filosofia da Linguagem deixaram, especialmente quando se tratava da determinação de significados de entidades abstratas. Os POMs e POTs, portanto, agregam conhecimento sobre metas abstratas do aparelho cognitivo humano (p. ex: um POM pode representar a meta "satisfazer as necessidades fisiológicas de um homem com fome" e um POT representar os quadros mentais desse homem ao se deparar com um restaurante, uma *pizzaria* ou uma churrascaria).

Vistas as teorias sobre conceito, chegou-se ao ponto de focar naquela que foi tributária de todas as linhas filosóficas e semânticas que se apoiaram na Lógica e na Matemática: a Teoria do Conceito da Terminologia, por várias vezes já citada ao longo do texto. Dos ensaios introdutórios, passa-se a estudá-la nos seus aspectos relevantes para a pesquisa

Todas as três acepções de conceito da Terminologia, explanadas no subitem 3.2.2.2.1, terão reflexos nos resultados desta pesquisa, cujo produto, apesar de não ser de natureza terminológica, emprega parte da metodologia desse campo para obtê-lo.

SILVA (2001, p.11) relaciona alguns motivos que mostram a importância da Terminologia, dos quais foram selecionados os que mais se aproximam do objetivo geral desta pesquisa:

- Ordenação do conhecimento;
- Transferência de conhecimento;
- Armazenamento e recuperação da informação técnica e científica.

Vale a pena citar o objetivo de um trabalho terminológico, para dar ainda mais relevo à importância da Terminologia para o marco teórico desta tese. Esse objetivo é a fixação de conceitos, para que sejam criadas definições e estabelecidos princípios para a criação de novos termos. Tais ações permitem uma comunicação mais precisa entre especialistas de várias áreas do conhecimento, no âmbito da Ciência e da Tecnologia (SILVA, 2001, p.12).

E o **termo**, esta unidade de trabalho básica para a Terminologia? Já se adiantou uma breve descrição, atribuindo-se-lhe a capacidade de designação de um conceito por meio de uma unidade lingüística. Mas esta descrição é insuficiente para os objetivos específicos desta tese.

De acordo com FELBER (1984), o termo é um símbolo lingüístico, atribuído a um ou mais conceitos, com base em conceitos vizinhos. Ele pode ser uma palavra ou um grupo de palavras, uma letra, um símbolo gráfico, uma abreviatura, uma sigla e outras expressões ou fórmulas de base lingüística.

No jargão cartográfico, em particular, denomina-se ***toponímia*** ao conjunto de termos (nomes) dos acidentes naturais e artificiais do terreno.

Como os conceitos existem independentemente dos termos e não podem ser comunicados diretamente, necessitam de um meio, ao qual Frege[161] denominou *de "nomina appellativa"*, ou termo conceitual, sensível aos sentidos.

Os termos normalmente são designados por especialistas que se apropriam de palavras para determiná-lo. A conexão termo-conceito é efetuada de uma forma deliberada, em contraste com a palavra, cuja forma e conteúdo surgem inconscientemente. O termo, como representação do conceito, é dependente do sistema conceitual ao qual pertence. Portanto, a correspondência biunívoca entre ambos é relativa (SILVA, 2001, p. 14).

Do exposto, verifica-se que a Terminologia, recortando da Lingüística aspectos da Semântica e da Morfologia, trabalha essencialmente com o conceito e este não tem sentido isoladamente, por sua própria natureza.

Já é chegado o momento de definir mais precisamente *denotação* e *conotação*, para dar continuidade lógica à apresentação do trabalho. Enfoques da Lingüística e da Terminologia serão examinados.

---

[161] Em sua carta de 25. Ago. 1900 para H. Liebmann [*apud* ALCOFORADO (1978, p.155)].

Segundo PINTO (1977), na Lingüística, a **denotação** é a relação entre uma expressão lingüística (termo) e uma classe de entidades à qual ela se aplica. É uma relação externa ao significado (conceito) lingüístico.

No meio lógico e lingüístico, às vezes se confunde denotação com **extensão**, mas PINTO (1977) coloca uma sutil diferenciação nas definições dos dois conceitos. A segunda é uma relação que se dá entre uma expressão lingüística (termo) e uma entidade ou classe de entidades *empíricas* que a relação indica. Também é uma relação externa ao significado (conceito) lingüístico. Como disse o autor: "Toda expressão possui denotação, mas nem sempre as expressões possuem extensão[162]".

Para o mesmo autor de antes, a **conotação** é uma relação que ocorre entre a expressão lingüística (termo) e o conjunto de características cuja posse pela entidade é a condição necessária para que a expressão a ela se aplique.

De acordo com essa visão lingüística, as características conotativas definem a denotação da expressão, ou de forma mais lógica: As características conotativas {ser humano, adulto, do sexo feminino}, p.ex., representam a condição necessária e suficiente para que o termo (expressão) "mulher" se aplique (defina) à classe de entidades femininas {Ana, Patrícia, Lívia, etc.} do mundo, i.e., esta classe de mulheres é denotada pelo termo conceitual "mulher" (*"nomina appellativa"* de Frege).

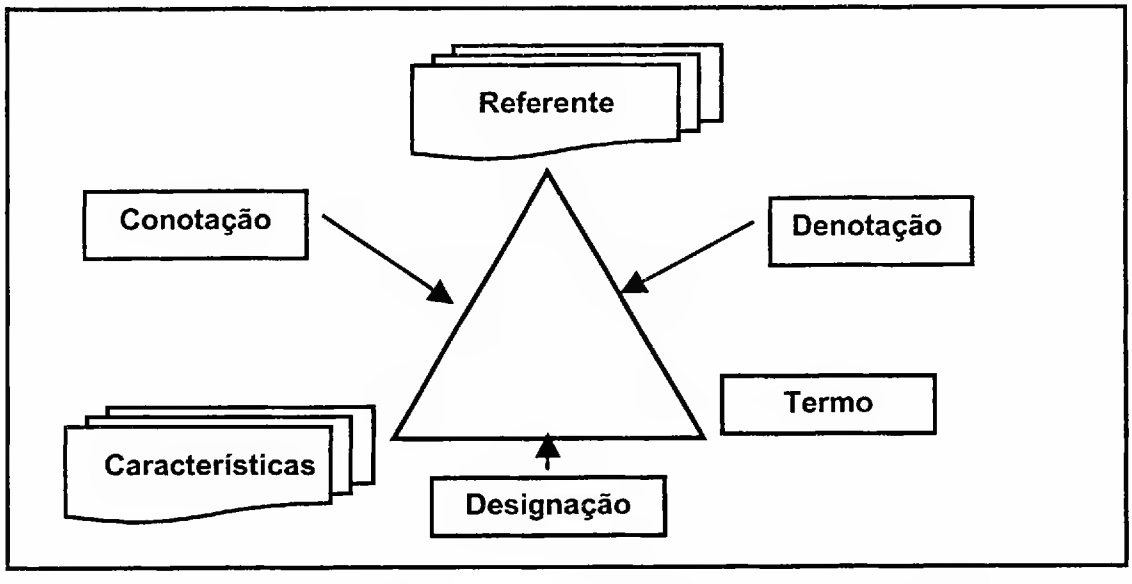

Figura 3.7: A tríade terminológica para o conceito.[163]

[162] Um unicórnio, p.ex.
[163] Adaptada de GOMES (1978).

Segundo ALCOFORADO (1978, p.155), Frege salientava esta função mediadora do conceito, quando escreveu o seguinte trecho numa carta para um colega seu:

"Inicialmente, devo enfatizar a profunda diferença que se dá entre conceitos e objetos, que é a de que nunca um conceito pode substituir um objeto, nem um objeto, um conceito ... A essência dos conceitos pode ser caracterizada pelo fato de se dizer que têm uma natureza predicativa. Um objeto nunca pode ser predicado de algo... Lingüisticamente, os termos conceituais ("nomina appellativa") correspondem aos conceitos."

GOMES (1978) ilustrou na Figura 3.7 o esquema triangular de DAHLBERG (1978), a fim de facilitar o entendimento da noção de conceito no âmbito da Terminologia.

O triângulo da Figura 3.7 pode ser entendido como o conjunto das propriedades significativas atribuídas a um referente pelos membros de uma área de assunto e sintetizado num signo lingüístico, que é o **termo** (GOMES, 1998). Essa esquematização parece ter tido sua origem nas definições de denotação e de conotação que vieram da Lingüística, apenas se diferenciando pela inserção de uma relação de *designação*, interposta entre as características da classe de entidades e o termo por elas conotado.

Concluindo, conceito, pela visão terminológica, é uma representação mental de objetos individuais. Um conceito pode representar somente um objeto ou – pela abstração – ser composto de um conjunto de objetos individuais tendo certas qualidades em comum e serve como um significado para a ordenação mental (classificação), com a ajuda do símbolo lingüístico (termo, letra, símbolo gráfico). O conceito é, assim, um elemento do pensamento ou uma representação mental não somente de seres ou coisas (expressos pelos substantivos), mas, num sentido mais amplo, de ações (expressas por verbos ou verbos substantivados) e também de localizações, situações ou relações (expressas pelos advérbios, preposições, conjunções ou nomes) (FELBER, 1984, p.115).

E as ***relações*** que existem ***entre conceitos***?

Segundo SILVA (2001, p.27-34), os tipos de relações são citados com poucas diferenças entre os autores, ocorrendo apenas opiniões diferentes quanto à classificação das relações.

As relações ***lógicas*** e ***ontológicas*** são conhecidas desde Aristóteles. Foi ele quem primeiro fez a diferença entre esses dois tipos de associações de idéias. As lógicas, baseadas em relações de equivalência (similaridade, identidade e outras). As ontológicas, baseadas na contigüidade ou na justaposição no tempo e no espaço (FELBER, 1984, p.102).

As relações lógicas têm recebido também o nome de relações genéricas ou relações de semelhança. As relações **lógicas** são as seguintes (FELBER, 1984, p. 190):

- Relação lógica de **identidade**: a relação de identidade indica que todas as características dos conceitos são idênticas. Por exemplo: **A** = lei e **B** = "law". Os conceitos **A** e **B** são idênticos, apesar de serem denotados por termos diferentes.
- Relação lógica de **implicação**: também conhecida como relação entre **gênero-espécie**, em que a espécie determina o gênero, ou seja, uma vez conhecido um termo específico e seu gênero, relacionados entre si pela implicação, é possível deduzir que as extensões do termo específico estão incluídas no gênero, isto é, o termo específico é um subconjunto do termo genérico. É curioso notar que o termo específico possui *intensão* maior que a do genérico e, dessa forma, assume todas as características do termo genérico, além de suas próprias características peculiares.
- Relação lógica de **interseção**: ocorre quando dois conceitos apresentam características comuns e não estão relacionados pela implicação (gênero-espécie); p.ex: arma e punhal.
- Relação lógica de **disjunção**: apresenta-se quando dois ou mais conceitos estão relacionados ao mesmo conceito, num mesmo nível de subordinação, e se excluem entre si, ou seja, uma extensão de um conceito não pode pertencer também ao outro conceito; p.ex: os conceitos de rapaz e moça, relacionados ao mesmo conceito de adolescente.
- Relação lógica de **negação** ou oposição: dois conceitos são opostos quando um conceito contém uma característica cuja negação existe como característica do outro; p.ex: compressão e descompressão.

As relações lógicas não podem cobrir todas as necessidades de ordenação conceitual em trabalhos terminológicos. Dessa forma, outras relações devem ser estudadas: as **ontológicas** (FELBER, 1984, p.198).

As relações ontológicas, isto é, as relações entre os seres, no caso, as entidades referentes, baseiam-se na contigüidade ou no contacto - no espaço e no tempo - dos elementos que formam o sistema conceitual para representar essas entidades. Por conseguinte, as relações ontológicas dividem-se em relações conceituais no espaço e no tempo ou relações **ontológicas verticais partitivas** no espaço e no tempo. Essas relações partitivas também são chamadas de relações do tipo **todo-parte** (FELBER, 1984).

Afim de **harmonizar** as pequenas **discrepâncias terminológicas** ou de nomenclatura convencional, as relações lógicas e ontológicas de FELBER (1984) que são mais manipuladas noutros campos do conhecimento são assim denominadas:

- Na **OO** (FURLAN, 1998): as lógicas de *implicação* são denominadas de relações de *generalização-especialização* e as ontológicas de *todo-parte* no espaço, de relações de *agregação regular* ou de *composição* (agregação por valor). No caso das classes de obje-

tos que estão num nível maior de generalidade, a denominação é de *superclasse*. No caso das classes mais especializadas, de *subclasses*.

- Na **Lingüística Computacional** (CLUL, 2002): as lógicas de implicação são denominadas de relações de *generalização*, em que os termos *superordenados* são denominados de *hiperônimos* e os subordinados, de *hipônimos*. As ontológicas de todo-parte no espaço são denominadas de *meronímicas*, em que os termos que se decompõem ou que agregam os componentes são chamados de *holônimos* e os que são componentes ou que se agregam aos primeiros (o composto) são chamados de *merônimos*.

Ao longo deste texto, a nomenclatura predileta é a da área da Lingüística Computacional. No entanto, quando se tratar de assunto mais específico da OO, a nomenclatura tenderá para a deste domínio científico.

As relações todo-parte no espaço são aquelas nas quais objetos complexos são constituídos de partes agrupadas, formando um todo, que muitas vezes têm características completamente novas, diferentes das suas partes elementares; p.ex: "carro" sendo o holônimo de "motor", "chassi" e "carroçaria".

Diferentemente, a relação todo-parte no tempo admite que as partes sejam reunidas ao conjunto em diferentes fases do tempo. As relações todo-parte no tempo têm sido utilizadas com muito êxito para processos de fabricação, trâmites alfandegários, seqüências evolutivas no reino animal e assim por diante. Nessa relação, as fases de um processo são descritas como partes de um todo (FELBER, 1984, p.201).

Segundo FELBER (1984), há outras relações ontológicas para atender a outros ramos do saber e o próprio campo de pesquisa sobre aplicações de novas relações desse tipo permanece em aberto para futuros estudos e sistematizações.

É chegado o momento de "equacionar" o conceito. Esta tarefa só pode ser realizada se o conceito passar por um processo de definição (um dos *modus sciendi*).

Muito se tem escrito sobre a definição da "**definição**". Para FELBER (1984, p. 178), sem dúvida, o denominador comum de todos os tipos de definições é expressar algo sobre um dado extralingüístico por meio de palavras, de termos ou de signos lingüísticos. Este mesmo autor se baseia numa norma alemã (DIN 2330) para formalizar o que seja *definição*:

"A definição é uma fixação de um conceito, com o fito de estabelecer relações com outros conceitos (conhecidos e já definidos) e de delimitá-lo em relação aos outros conceitos."

Dessa definição, podem ser deduzidas três funções, segundo o mesmo autor citado:

- Fixar um conceito (freqüentemente mediante sua normalização, após prévio acordo);
- Delimitar um conceito (isolá-lo de outros conceitos); e

- Relacionar um conceito com outros conceitos (o que leva, conseqüentemente, à estruturação de um sistema de conceitos).

A estrutura formal da definição tem muita semelhança com uma equação, como a que a formulação a seguir ilustra:

$$T = G + d_1 + d_2 + ... + d_n$$

Esta expressão foi denominada de *estrutura formal da definição predicativa*[164] por GARCIA (1976, p.308), mas também aparece noutras formas em FELBER (1984, p.179) e RODRÍGUEZ (2000, p.107).

Os seus elementos são assim especificados:

- T (1° termo da expressão): *definiendum* (o que se quer definir);
- 2° termo da expressão: *definiens* (o que define - *características*);
- = : cópula;
- G: gênero (classe à qual o termo imediatamente se subordina);
- d_1 ... d_n : diferenças (o que descreve - *propriedades*)

Relembrando Frege, acompanhado por GORSKI (1968) e GARCIA (1976), só se definem as classes de entidades do mundo pelas *características*; as *propriedades* descrevem as instâncias dessas classes.

As características também dispõem de funções assessórias, como:

- Validar o conteúdo semântico do conceitos;
- Detectar alterações da *intensão* de um significado muito estrito, que acabam transformando-o num novo conceito;
- Elaborar e formular definições;
- Estruturar os conceitos dentro de um sistema.

É interessante notar que a estrutura formal da definição relaciona dois conceitos, o *definiendum* e o *definiens*, por meio de uma relação de implicação (gênero-espécie). O *definiens* é um termo genérico mais próximo e o *definiendum* é um termo específico (DAHLBERG, 1978, p.149). Essa estrutura formal foi adaptada para a tese de RODRÍGUEZ (2000, p. 47 e 107) numa estrutura que foi a base para a montagem das ontologias que seriam carregadas em seu modelo de avaliação de SS (MSS). A Tabela 3.1 fornece uma idéia de como serão montadas as ontologias no Capítulo 6 (metodologia).

---

[164] O A. usou "denotativa", mas optou-se por esta alternativa do mesmo autor, para guardar harmonia com a terminologia de outros autores da área [(FELBER (1984) e PINTO (1977)], que distinguem *denotativo* de *conotativo* ou *predicativo*.

Do exposto, *definir* é relacionar conceitos por meio da relação de gênero-espécie e estabelecer as características próprias do termo específico.

Além da definição clássica estudada, que DAHLBERG (1981, p. 17) chama de conceitual (gênero próximo + propriedades específicas), e que se fundamenta na relação entre os três componentes do conceito - *características*, *referente* e *termo* -, existem outros tipos de definição (FELBER, 1984, p.179):

- Definição específica;
- Definição genérica;
- Definição por exemplo.

Tabela 3.1: Componentes da representação de uma classe de entidades.

Componentes			**Descrição**
Definiendum			Conjunto de sinônimos que se referem à classe de entidades.
Definiens			Termos utilizados para definir a classe de entidades.
	Relações semânticas		Relações verticais entre uma classe e outras classes
		Gênero-espécie (lógica)	Relações do tipo *é-um(a)*
		Meronímicas (ontológica)	Relações do tipo *parte-de*
	Feições distintivas		Propriedades da classe de entidades
		Partes	Elementos estruturais
		Funções	O que as instâncias das classes fazem ou o que é feito nelas.
		Atributos	Propriedades que não se enquadram nem como partes nem como funções.

A **definição específica** ou **definição por *intensão*** é aquela determinada pela compreensão (***intensão***), i.e., parte-se do gênero imediato e das características restritivas que diferenciam este conceito dos que lhe são próximos, no mesmo nível de abstração. Nessa definição, todas as características detectáveis do conceito são usadas, provocando, indubitavelmente, confusão. Não é aconselhável o seu uso.

A **definição genérica** ou **definição por extensão** é aquela determinada pela ***extensão*** do conceito, i.e., enumeram-se todas as espécies de um gênero que estão no mesmo nível de abstração. Normalmente, são de três tipos:

- Enumeram-se todos os *objetos individuais* que pertencem ao conceito, p.ex, "Os planetas do sistema solar são: Mercúrio, Vênus, Terra, Marte, Júpiter, Saturno, Urano, Netuno e Plutão". Esse tipo de definição exige que a enumeração seja absolutamente completa e a menor alteração acarreta a retificação da definição.

- Enumeram-se, no *mesmo nível de abstração*, todos os conceitos subordinados ao conceito a definir, p.ex, "Nesta bacia, são tipos de cursos d' água: arroios, regatos e riachos". Esse tipo de definição é mais resistente à mudança que a anterior, visto que é mais difícil aparecer uma classe nova de entidades do que uma entidade individual.
- Indica-se a regra pela qual se chega a uma enumeração, p.ex, "Número primo é aquele que somente é divisível por ele mesmo ou pela unidade".

A **definição por exemplo** é aquela em que o termo a definir se apresenta como uma frase, cujo significado completo se conhece ou pode estimar-se (FELBER, 1984, p.181). Esta definição deve ser evitada para uso terminológico, porque é muito imprecisa.

E as **relações** que existem **entre** os **conceitos** e os **termos** ?

Em primeiro lugar, é preciso diferenciar um *termo* de uma *palavra*.

O ***termo***, como já definido, é uma simbolização gráfica de um conceito, podendo assumir a forma de uma palavra ou mesmo de um grupo de palavras. O termo pertence a um sistema terminológico, i.e., cumpre uma finalidade classificatória do conhecimento de uma certa língua profissional (LP) e se caracteriza pela precisão.

A ***palavra*** já não se insere num sistema tão preciso como o termo, estando no domínio mais amplo da língua geral e não constitui o cerne de uma definição científica, apesar de ter a possibilidade de também compô-la (FELBER, 1984, p.213).

Às vezes, para revestir a palavra da precisão necessária e transformá-la num termo, é preciso contar com meios extras numa definição, que explicam de modo mais rápido e com mais exatidão o que se trata (FELBER, 1984, p. 185). Contudo, não é possível substituir uma definição verbal por estes exemplos de meios extras: Ilustrações, mapas, esboços, exemplos e fórmulas (exceto se utilizadas entre especialistas).

Ao lado de FELBER (1984), o cientista cognitivo David Mark (MARK, 2002) também recomenda a definição e não os subsídios gráficos para trabalhar de forma o menos ambígua possível com um sistema conceitual, conforme se verá no subitem 6.3. Para ambos, tais formas de representação são abstrações que às vezes fogem aos padrões de objetividade do sistema conceitual em questão

Um dos objetivos de uma LP é produzir definições sem o efeito indesejável da ambigüidade. Para alcançar esse fim num estudo terminológico, é preciso usar termos o mais precisamente possível e examinar as relações entre os conceitos e os termos que os representam, subdividindo-as em cinco tipos: *monossemia*, *polissemia*, *sinonímia*, *equivalência* e *homonímia*.

A ***monossemia*** ocorre quando um termo só designa um conceito. Esta é a situação ótima em matéria terminológica, mas que ocorre raramente, somente se um esforço de normalização for empreendido.

A ***polissemia*** se dá quando um termo designa dois ou más conceitos distintos, mesmo assim guardando certa semelhança entre si e mesmo que não pertençam ao mesmo sistema de conceitos; p.ex: o termo "atração", no contexto artístico, pode designar um espetáculo; no contexto da Física clássica, designar a força gravitacional entre dois corpos e, no contexto da Biologia, designar a necessidade de manter contacto sexual entre dois seres de sexos opostos. Apesar do aspecto relevante da polissemia no processo de criação de novos termos numa LP, ela nem sempre é um fator positivo para a comunicação entre dois ou mais usuários dessa espécie de língua. Confunde-se muito polissemia com homonímia (a seguir), mas a diferença entre essas relações, apesar de sutil, existe.

A **homonímia** (mesmo significante e significados distintos) ocorre quando um termo designa dois ou mais conceitos entre os quais não existe nenhuma relação semântica; p.ex: acender (iluminar) x ascender (subir) e cegar (tornar cego, sem visão) x segar (ceifar) são homônimos *homófonos*; leste (l"ê"ste – forma passada de ler) x leste (L"é"ste – ponto cardeal) e sede (s"ê"de – necessidade de água) x sede (s"é"de – local) são homônimos *homógrafos*. A homonímia, em geral, não constitui um grande óbice comunicativo.

A **sinonímia** (mesmo significado e significantes distintos) ocorre quando dois ou mais termos da mesma língua designam exatamente o mesmo conceito. Em geral, esta relação entre termos e conceitos é prejudicial à comunicação por sugerir diferenças que não existem, derrubando princípios básicos da Terminologia: claridade e simplicidade na comunicação. Causas da sinonímia:

- A falta de conhecimento sobre o assunto;
- Normalização fracassada;
- Criações de termos *ad-hoc* por tradutores que não pertencem ao campo do saber.

A ***equivalência*** é uma relação muito parecida com a sinonímia, só que ocorre entre fenômenos de sistemas lingüísticos diferentes.

Conhecer as atividades típicas de um trabalho terminológico é uma forma de aprimorar as construções de definições das classes de entidades espaciais.

SÖRGEL (1999) já reclamava da necessidade dessa sinergia, que SILVA (2001) explorou em seu trabalho, comparando linguagens de modelagem da Ciência da Computação (UML™) e da Terminologia (*tesauro*), tendo verificado que ambas se aproximam pela metodologia de modelar a realidade da forma mais simples e clara possível.

Algumas das atividades típicas de um trabalho terminológico vêm na seqüência:

- Selecionar e registrar os termos correspondentes aos conceitos de um determinado campo do conhecimento;
- Descobrir, criar e normalizar um sistema de conceitos sobre um assunto determinado;
- Descobrir e normalizar as correspondências entre conceito e termo;
- Descrever conceitos por meio de explicações e definições ou normalizar as definições existentes (diagrama de conceitos).

Uma ferramenta primordial para elaborar um sistema conceitual é **o diagrama de conceitos**, que deve ser o mais inequívoco possível. Essa precisão é alcançada quando se representam as relações lógicas e ontológicas entre conceitos de um ramo específico do saber. FELBER (1984, p. 229) colocou como pré-requisito para a elaboração de um trabalho terminológico a credencial do especialista (ou equipe) de uma área do conhecimento humano, com formação ou capacitação em Terminologia

FELBER (1984, p. 69) reportou que, na década de 60, organismos europeus, preocupados com a Terminologia, começaram a automatizar determinados procedimentos de armazenamento e de busca de dados terminológicos. Provavelmente esses esforços culminaram no catálogo SDTS™ (*Spatial Data Transfer Standard*) e na taxinomia *on-line Wordnet*™, que foram utilizados por RODRÍGUEZ (2000) para a construção de sua ontologia. Na verdade, o catálogo, o vocabulário e a ontologia desta última autora constituem exemplos de diagramas conceituais.

A seguir, uma discriminação desses dados terminológicos:

- Símbolos lingüísticos, que representam os conceitos em forma de termos;
- Descrições lingüísticas, que descrevem a *intensão* (conteúdo) dos conceitos em forma de características, mediante uma definição;
- Símbolos ou convenções gráficas, que indicam as relações de um conceito com os restantes de um campo conceitual dado (relações lógicas e ontológicas).

Esmiuçando ainda mais os dados terminológicos, cada um dos três tipos acima citados pode ainda ser decomposto da seguinte forma:

- Símbolos lingüísticos: termos e sinônimos que se destinam às especificações de normalização;
- Descrições lingüísticas de conceitos: definições, fórmulas e contexto de definição;
- Relações: as já citadas de gênero-espécie e as meronímicas, particularmente.

FELBER (1984, p. 223) também levantou algumas recomendações fundamentais para a construção de termos, que muito serão aproveitadas para as definições necessárias à

constituição dos dois instrumentos de pesquisa descritos no Capítulo 6 desta tese: o questionário e o PRONTO©. A seguir, as recomendações mais relevantes:

- Um termo deve ser lógico e sumamente auto-explicativo;
- Um termo é mais lógico e sumamente auto-explicativo na ordem direta do conhecimento técnico-profissional de seu criador;
- Um termo deve ser sistemático; p.ex: tesoura para cortar papel (um termo) e tesoura para aparar grama (outro termo);
- A criação de termos deve obedecer às regras gramaticais da língua em questão;
- Um termo deve dispor de muitas possibilidades de derivação; p.ex: telefone, telefonia, telefonista, e assim por diante;
- Um termo não deve ser pleonástico ou redundante; p.ex: escala graduada (uma escala já traz em si uma graduação);
- Um desdobramento da recomendação anterior está no fato de que um termo não deve conter elementos supérfluos; p.ex: mineral de quartzo (o quartzo já é um mineral);
- Um termo deve ser o mais curto possível; p.ex: máquina acepilhadora no lugar de "máquina de aplainar madeira";
- Um termo, se possível, não deve ter homônimos, nem sinônimos e nem ser polissêmico.

Por ser desejável que um termo não tenha sinônimos, não significa que estes não possam ser usados para estimar a SS entre dois conceitos pela pontuação alcançada entre as coincidências dos termos sinônimos, com respeitos às suas características e propriedades. Este recurso será utilizado no PRONTO© com a denominação de *consin* (conjunto de sinônimos).

Mais adiante, em BÄHR (1996), essas noções encontraram uma proveitosa aplicação na rede semântica da Figura 3.9. O referente "árvore" daquela rede é o objeto formal ou *constructo*. A árvore não existe fisicamente. É uma abstração. O que existe na realidade é uma determinada árvore com certas características, outra árvore, com outras características, enfim, uma coleção de objetos que se enquadram no conceito de árvore. O termo é o signo lingüístico que sintetiza o conjunto de características significativas atribuídas à classe "árvore". A relação entre o termo "árvore" e o referente não é direta. Ela é feita por intermédio do conceito. É preciso existir **referente**, **características** e **termo** para haver **conceito**. Analisar o *referente* "árvore" é fazer predicações (conotações, *intensões*) verdadeiras, i.e., identificar e descrever as suas propriedades. Definir o *conceito* de árvore é identificar as propriedades do referente "árvore" que permitem relacionar o conceito de árvore a outros conceitos, o que propicia a construção de um sistema de conceitos (GOMES, 1998).

As presentes considerações podem ser muito promissoras na avaliação de certos fatores que entram na construção de uma rede semântica ou de uma **taxinomia**. O aporte teórico que esta pesquisa necessita dessa área não se prende tanto ao binômio *saussureano* significante-significado, mais vazado na Lingüística, porém num processo mais abrangente, que garanta o uso interpessoal da língua para a comunicação.

O que mais interessa do trabalho de Saussure, no presente contexto, foi a sua convicção da existência de um *invariante* em nível mais profundo, comum a todos os que se utilizassem de um sistema de códigos lingüísticos para comunicação. Assim, não apenas as ciências sociais tirariam proveito desse núcleo estável de características sistêmicas, capaz de passar por um processo de formalização, mas as naturais e, especialmente, as *geociências*, que também lidam com a comunicação.

CÂMARA (1999), em reforço a BÄHR (1996), pôs em relevo a aquisição e a representação de objetos espaciais (*geoespaciais*), ao analisar as limitações dos métodos matemáticos e estatísticos tradicionais, que não consideram o fato de o fenômeno expresso pela superfície física terrestre variar de forma diferente e em direções distintas no espaço (anisotropia). Assim, o autor exaltou as vantagens da Lógica *Fuzzy* (V. glossário) em relação à Lógica tradicional. A Lógica *Fuzzy* propiciaria a construção de uma grade numérica (matriz) que seria associada a uma "superfície de decisão", adequando-se mais fielmente à variação contínua do fenômeno no mundo real. São formas alternativas e também formais de explorar aquele núcleo lingüístico e estável de Saussure, de natureza sistêmica, que abriga (porta) o significado.

MARTINELLI (1991) expôs de forma clara, simples e concisa os fundamentos teóricos da obra de BERTIN (1967). O autor estudou as três relações fundamentais entre objetos espaciais: a *diversidade*, a *ordem* e a *proporcionalidade*, ao tentar fazer da representação gráfica uma linguagem para a comunicação humana, livre da ambigüidade que afeta a língua natural. Nessa obra, o autor aplicou esses conceitos à Cartografia, prescrevendo regras simples para a elaboração de mapas temáticos e mesmo cartas topográficas, de forma a economizar tempo, recursos e garantir uma expressão visual clara e descongestionada de pormenores insignificantes para a vista do usuário.

MOURA (1994) ressaltou a necessidade do estudo de Semiologia Gráfica, para prover fundamentação teórica à Cartografia como veículo de comunicação. "A Semiologia Gráfica transcodificaria a linguagem escrita para a gráfica, evitando o ruído na comunicação." A busca por signos que representem fielmente as características mapeadas estaria em grau de

correlação direta com as propriedades de percepção visual e nas características dos sistemas em que os sinais acumulam significados, concluiu MOURA (1994).

Um trabalho que seguramente deu impulso a esta pesquisa foi o de MEDEIROS (1999, p. 252, 254 e 259). A autora [assim como PINTO (1977)] incentivou a continuidade de estudos sobre a análise semântica de *corpora* de outros campos do conhecimento, ao indicar a sua metodologia, especialmente na organização dos conceitos pertinentes a campos específicos de estudo e na montagem das ontologias correspondentes.

O objetivo geral da pesquisa de MEDEIROS (1999) foi o de solucionar casos de ambigüidades[165] em textos científicos e técnicos de uma LN (no caso, o português). O seu foco foram elementos frasais, constituídos de sujeito, predicado e complemento, o que afasta os marcos teóricos do trabalho dessa autora do da presente pesquisa, que só foca os termos, distantes de uma estrutura frasal e sem preocupação com a avaliação de ambigüidade desses termos e do contexto em que se inserem.

Mesmo assim, MEDEIROS (1999) e RODRÍGUEZ (2000) coincidem em inúmeros pontos, particularmente quando admitem que o contexto é um fator de controle de formas polissêmicas de uma LN (no caso de estruturas frasais, para a primeira autora) e de uma língua profissional (no caso de termos, para a segunda autora). Ambas trabalharam com esse complexo fenômeno (contexto) e obtiveram resultados significativos. Suas ontologias, apesar de distintas pelo conteúdo dos *corpora* utilizados, serviram para o mesmo objetivo: passaram a estrutura conceitual que representavam pelo crivo de um instrumento de análise semântica - o Zstation™ para a primeira autora e o MSS para a segunda autora.

Apesar de ostensivamente ter sido declarado ao longo desta obra que o contexto não foi utilizado na avaliação da similaridade semântica (que não deixa de ser uma análise semântica ou *componencial*), a manutenção de certo controle sobre variáveis intervenientes pressupõe a existência de um contexto no campo de observações da folha Faxinal, alvo deste estudo-de-caso. As variáveis que foram mantidas sob controle foram as que não o foram, intencionalmente, nos trabalhos das outras duas autoras citadas, visto como ambas necessitavam do relaxamento desse controle, para poder detectar as influências no valor do significado das expressões de seus *corpora*.

RODRÍGUEZ (2000), p.ex., utilizou respondentes de um curso de graduação em Letras para os questionários de cunho *geocientífico* que formulou.

Nesta tese, não. Para manter um contexto homogêneo, i.e., que não afetasse de forma salientada alguma feição distintiva (*parte, função* ou *atributo*, como se verá adiante) das

classes de entidades espaciais que participaram da manipulação experimental, os respondentes foram selecionados de um mesmo ambiente de trabalho, que dominam a mesma LN e a mesma língua profissional (*geoprocessamento*), que se formaram nos mesmos estabelecimentos de ensino, enfim, indivíduos que partilham do mesmo espaço de estimulação cognitiva. Tal escolha, sem dúvida, manteve relativamente constantes efeitos subjacentes, provocados por um contexto (ambiente) de natureza heterogênea.

A diferença metodológica entre esta pesquisa e a de MEDEIROS (1999) é que esta autora utilizou um instrumento já pronto para apoiar o seu teste de hipótese. Esse instrumento foi um SE - o **Zstation**™, utilizado no estudo da **ambigüidade** que afeta a **forma verbal** dos signos. Nesta pesquisa, todavia, foi desenvolvido um instrumento *ad-hoc* para apoiar o teste de hipótese, na **avaliação da similaridade semântica** entre os termos armazenados numa estrutura de conhecimento, na forma de uma rede semântica (taxinomia). Esse instrumento foi denominado de **PRONTO**© e foi codificado em *Java*™.

Vale frisar que entre os trabalhos revisados em que se buscava aplicar teorias ligadas a signos lingüísticos, na sua acepção mais ampla (Semiologia de Peirce), i.e., incluindo os signos não-verbais de Saussure, PRADO (2000) foi o autor que obteve mais sucesso, quantificando os fenômenos geográficos à luz das suas semelhanças e diferenças semânticas.

O objetivo da análise comparativa de PRADO (2000) entre SIGs e sistemas semióticos[166] foi o de levantar requisitos para interfaces mais amigáveis com usuários gerais de tecnologias de *geoprocessamento*. Ele partiu da premissa de que seria um desperdício um usuário não especializado em SIG, não tirar vantagem da inigualável capacidade desses sistemas para produzir conhecimento, simplesmente por não haver uma forma de interagir com as estruturas de comunicação desses sistemas.

PRADO (2000) estudou três autores preocupados com representações visuais em Cartografia: 1) O já citado Jacques Bertin, 2) Raul J. Ramirez e 3) Ian Pratt. Dos três, aproveitou o que de melhor havia para construir o seu teste. O critério de escolha desses trabalhos foi justamente a preocupação dos autores com os processos de *percepção* e de *interpretação* das representações visuais. A seguir, um resumo do que produziu cada um dos 3 autores pesquisados por PRADO (2000).

J. Bertin coletou e analisou diversas simbologias gráficas e acabou por derivar um conjunto de sete variáveis visuais para mensurar os elementos gráficos pontuais, lineares e planares; são elas: *posição, tamanho, valor, granulação, cor, orientação* e *forma*. O problema

---

[165] Aplicou a Teoria Gramatical das Valências de Borba.
[166] Neste trabalho: mapas, imagens, fotografias...

desse enfoque é que Bertin limitou muito o seu estudo às características bidimensionais do suporte gráfico, bem como desconsiderou fenômenos de natureza cronológica.

Ramirez tentou orlar um *alfabeto cartográfico*, i.e., elementos básicos (primitivas) com os quais os mapas seriam construídos[167]. Assim, esse alfabeto seria formado de: {ponto, segmento de reta, segmento de curva e espaço vazio}. Quando criou suas "regras gramaticais" para construir mapas, chegou às mesmas variáveis visuais de Bertin. O problema com esse enfoque foi a excessiva *granularidade* a que chegaram as "sentenças" de Ramirez para construir seus mapas. Esse fator o levou a elementos desprovidos de significado, prejudicando o estudo da comunicação visual, que é um fenômeno sempre presente na construção ou na interpretação de um mapa.

Como Ramirez, Ian Pratt também admitiu que existe muito em comum entre mapas e outras formas de comunicação, como as línguas naturais. Ambas as formas de expressão utilizam marcas de tinta sobre o papel, por meio de um código preestabelecido. Entretanto, ressaltou que o que diferencia os mapas é a sua característica geométrica e visual, aproximando-o das entidades do MR que se deseja representar.

PRATT (2000) chegou mais perto da semântica cartográfica, ao obter um conjunto de regras de validação do conteúdo de um mapa. Para tanto, lançou mão da definição de Hansgeorg Schlichtmann, em que cada elemento cartográfico pode ser decomposto em duas partes: o *substantivo* (natureza da entidade = propriedades visuais; p.ex., casa: □) e a *localização* (coordenadas do objeto representante da entidade no mapa).

Para que o leitor de um mapa seja capaz de interpretar os pares (substantivo, localização), presentes num mapa, é preciso decodificar a realidade geográfica expressa por tais pares. Para modelar esse fenômeno, o autor utilizou duas funções: uma de interpretação simbólica $I_{(x)}$, que correlaciona cada elemento $\boldsymbol{X}$ (substantivo) do domínio cartográfico com o conjunto de entidades da realidade geográfica que se quer representar. A outra função é $\mu(x, y)$, que correlaciona os pontos da superfície (x, y) do mapa com posições do espaço.

A Figura 3.8 ilustra um mapa [domínio das funções [$I(X)$ e $\mu\ (x_c, y_c)$] que contém os objetos representativos das entidades do mundo-real (imagem para as duas funções).

O caso teórico [Figura 3.8(a)] mostra a perfeita relação entre as entidades do MR e seus representantes no mapa.

---

[167] Baseou-se na Teoria dos Níveis Lingüísticos de Chomsky.

Na Figura 3.8(b), pode-se verificar que as funções de decodificação da realidade não são bijetoras, mas apenas injetoras, i.e., ocorreu um caso de ambigüidade na representação, em que não existe um representante do mapa para uma entidade do MR.

Tanto o objetivo do estudo de Pratt como o desta pesquisa não são os de descrever a função *I(X)*, ligada às formas de representação (variáveis visuais de Bertin). Para ambos os objetivos, importa apenas que ela exista. Como disse RODRÍGUEZ (2000), esta linha de pesquisa não está interessada nos aspectos gráficos da IG, mas nos lingüísticos.

Os objetivos específicos desta pesquisa não devem focar a clareza das relações representante – representado. Devem ser aceitas como pressupostos já comprovados. A preocupação dos objetivos de pesquisa é a de verificar se a BD "esconde" essas relações, traduzidas pela função *I(X)*.

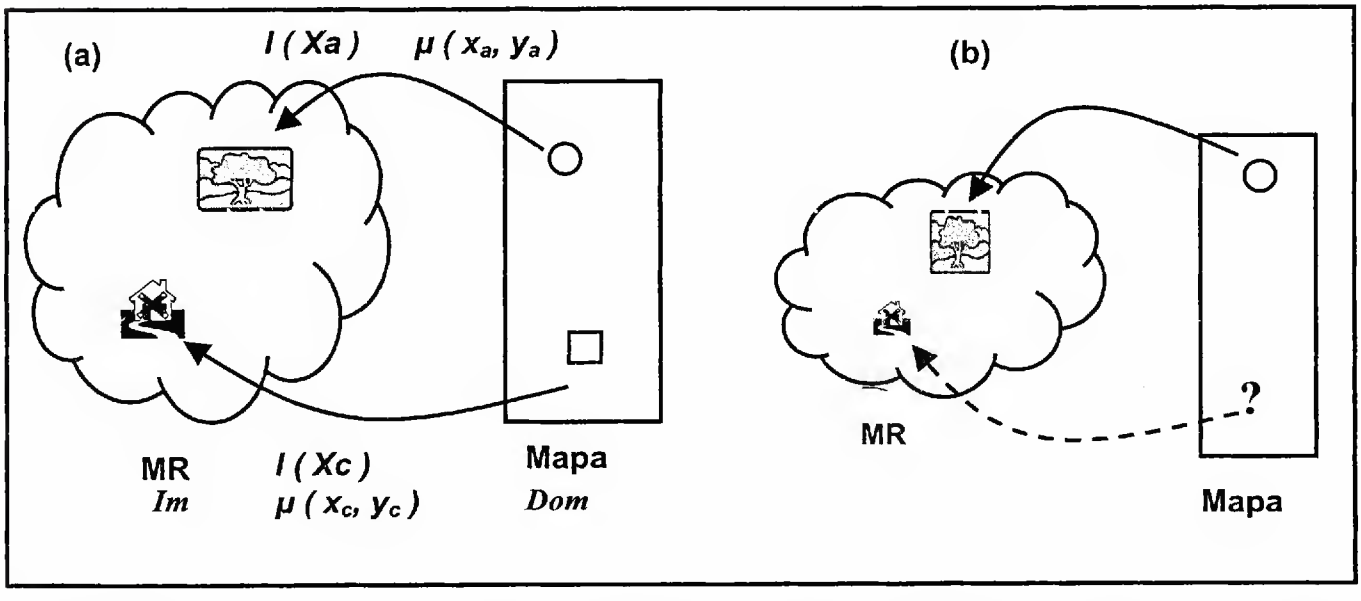

Figura 3.8: Função de Pratt para validação de um mapa: caso teórico ideal (a); presença de ambigüidade (b).

Na verdade, a limitação das interfaces dos SIGs tradicionais é um problema de comunicação homem-máquina (HCI) de duplo efeito danoso, porque além de não permitir a comunicação mais imediata, entre o homem e as funções do SIG (aquisição, análise e exibição da IG), interrompe também um outro processo de comunicação, num nível mais elevado, entre o conteúdo de um documento cartográfico e o aparelho cognitivo do usuário, que possui todas as potencialidades de modelagem da informação já descritas no subitem 1.5.3.1: percepção – indução – abstração.

As implicações de ordem metodológica que podem ser extraídas do trabalho de PRADO (2000) serão exploradas na seção dedicada à descrição da metodologia. Tais implicações envolvem questões como: "É conveniente reforçar as respostas aos questionários dos indivíduos com apresentação de documentos cartográficos? Ou é melhor que eles respondam com base apenas nas representações mentais que os termos denotativos das entidades espaciais provoquem?

Tais questões foram tratadas com RODRÍGUEZ (2002) por meio de correspondência e com um autor americano [(MARK, 2002), nos subitens p. 146 e no subitem 6.3], especializado no assunto e indicado pela autora. Depois das respostas obtidas, combinando-as com as conclusões do trabalho de PRADO (2000) e pelas informações obtidas numa recente palestra[168] sobre um procedimento inusitado de substituir as funções clássicas de banco de dados em nível conceitual (V. PREVAYLER™, no subitem 4.3), várias conclusões ainda poderão ser tiradas sobre a semântica da IG representada em bases de dados. Essas questões foram as que mais estorvaram a metodologia de desenvolvimento do PROFAX, mas que foram resolvidas satisfatoriamente no desenvolvimento do PRONTO©.

Até o parágrafo anterior, pode-se dizer que foram feitas análises da IG num nível de abstração simbólico (do signo). Daqui até o final deste subitem, serão tecidas considerações de ordem *subsimbólica*, i.e., num nível mais baixo de abstração, próximo aos *bits* e *bytes* e até num nível mais baixo ainda que este, no nível das grandezas físicas (luz, eletricidade), que sustentam todos os outros níveis superiores de informação (da comunicação lingüística). De certa forma, é uma extensão do subitem 3.2.1, com um viés semântico.

BÄHR (1996) defendeu que a aquisição e o processamento de dados espaciais apoiado por computador, requerem mais do que as já insatisfatórias tentativas de captura baseadas na modelagem geométrica. Explicou ser imprescindível a agregação da modelagem semântica a esse processo de captura, sob o risco de se ter à mão a formação de caras e volumosas bases de dados que, indiscutivelmente, deverão receber um tratamento complexo de conversão futura, para se adequarem às prescrições modernas de obtenção de informação "pura", i.e., sem ruído ou com o indesejável efeito da ambigüidade controlado.

BÄHR (1996) investigou as questões da ambigüidade da informação geográfica no seu aspecto mais genérico, pelo foco da modelagem semântica (campo da Lingüística) e sob o processamento de imagens (campo do *geoprocessamento* [169]). Isso caracteriza o que os autores chamam de "mudança de paradigma" numa disciplina ou área de estudos (tanto tec-

[168] Realizada em 15.08.02, na sede da Faculdade Alvorada (DF), por um grupo de pesquisadores de *Java*™ (DFJUG, 2002)
[169] V. glossário.

nológica como não-tecnológica), o que provoca uma profunda alteração de um sistema bem estrito de conceitos para formar um outro sistema que, apesar de se ter originado do anterior, não guarda com ele qualquer relação fundamental de dependência.

Os processos de aquisição[170] de dados *geoespaciais*, ao longo de menos de meio século, já passaram por 3 grandes fases [(LUGNANI, 1987) e (PRADO, 1992)]: *analógica* (baseada em modelos geométricos e nas medidas executadas em instrumentos óptico-eletro-mecânicos, sem apoio de computador); *analítica* (baseada em modelos algébrico-lineares transformados em algoritmos computacionais, que restringe ao mínimo possível a participação de instrumentos óptico-eletro-mecânicos nas medidas) e *digital* (baseada em modelos de correlação digital e automática de imagens, embutidos em estações de trabalho[171]).

O salto do modelo tridimensional geométrico, passando por formas de representação numérica e, agora, diante de modelos digitais que representam o terreno em níveis de cinza (resolução[172] espectral) por algoritmos computacionais que trabalham com funções estocásticas complexas, significa experimentar a mudança de paradigma científico apregoada por Thomas Kuhn. Aliás, nesse caso específico, não foi propriamente uma mudança da 1ª para a 2ª fase, mas sim da 1ª ou 2ª para a 3ª fase.

Apesar dos avanços tecnológicos introduzidos nos equipamentos de aquisição de dados espaciais da fase analógica, que eliminou muitas das rotinas braçais do operador humano – fonte de introdução de erros -, os métodos analíticos não extraíam (ainda não extraem) todo o conteúdo informativo das fotografias ou imagens da superfície terrestre. Com a introdução dos métodos digitais, as possibilidades que se abrem são inúmeras. Por outro lado, cada possibilidade já desvendada ou a desvendar traz consigo um germe de proliferação exponencial de problemas, constituindo um campo fertilíssimo de investigação científica.

A relação entre possibilidades e problemas exposta surge da natureza fractal e maleável ("plástica") do dado digital, baseada no estado binário (0 ou 1 – ausência ou presença de tensão elétrica), ou baseada no correspondente estado dicotômico da Lógica (verdadeiro ou falso). Esse estado binário simples é "digerível" numa velocidade vertiginosa e num alto grau de estruturação pelos circuitos digitais miniaturizados dos computadores modernos, o que já não ocorre com os circuitos analógicos, difíceis de construir e que "captam" o mundo real numa forma contínua e não fractal (sulcada, dividida).

---

[170] Para maiores minúcias, V. Fotogrametria no glossário.

[171] "*Workstation*", i.e., um sistema conjugado de *hardware* e *software*, altamente especializado, para executar tarefas complexas de projetos de engenharia, medicina ou qualquer outro campo avançado de conhecimento humano (BORGES, 1995).

[172] V. glossário.

Na forma contínua do fenômeno, é muito difícil e às vezes impossível isolar os ruídos que acompanham o sinal analógico (FITZGERALD, 1981). Além disso, o citado alto grau de estruturação a que estão sujeitos os dados digitais nos vários níveis de abstração de um computador digital (desde as portas lógicas – V. glossário -, impressas em circuitos até os *softwares* mais avançados), permitem a associação do fenômeno físico observado a um gradiente de estados, estabelecido segundo algum critério (arbitrário ou fundado cientificamente) do operador humano.

A estrutura digital do computador, que se adere à lógica *booleana* e tem poder de representação para esse gradiente de estados do mundo físico, possui uma representação matemática de emprego muito generalizado: o *vetor multidimensional* (*array* ou matriz) [(TANENBAUM, 1999), (NORTON, 1989)].

Em suma, um modelo geométrico (fase analógica) e, muito mais, um modelo numérico (fase analítica) da realidade podem falsear menos um fenômeno observado do que o modelo digital, somente se este último não for conhecido na sua essência. Se o modelo digital for bem compreendido e caracterizado em suas imensas possibilidades de representação (o que não existe no geométrico e no analítico), controlando-se os ruídos (ambigüidades) e adequando-se o fenômeno observado à melhor representação, nunca, na recente história dos processos automatizados de aquisição, processamento e reprodução da informação digital, ter-se-á atingido um grau tão alto de agregação de significado a esta informação (no caso, a *geoespacial* ou IG).

Fica mais fácil entender por que a informação transportada por um sinal digital carrega mais significado que no caso geométrico, p.ex., quando se compara um promontório (pico), que, para ser "captado" por um sistema analógico, é abstraído como sendo uma coleção de primitivas euclidianas (pontos, retas e polígonos fechados - áreas), que não existem no mundo real, mas só na mente humana, sendo os complexos mecanismos cognitivos de classificação, dedução e indução os responsáveis pela carga semântica atribuída ao promontório. Dessa maneira, nem a Matemática provê meios adequados de representar o *continuum* da realidade empírica e, muito menos, os recursos óptico-eletro-mecânicos da tecnologia analógica.

De outro modo, o mesmo promontório pode ser "captado", numa primeira fase (de campo), por um sistema sensor (óptico ou munido de RADAR – V. sensores no glossário) e, numa segunda fase (de gabinete), pelo mecanismo de aquisição de dados de um sistema de informações geográficas, de forma totalmente digital. Nesse ambiente computacional, a semântica que está por trás dos *bits* e *bytes* dos dados digitais adquiridos ("captados") pode

ser revelada ao usuário desses sistemas das formas mais versáteis que este o desejar, tendo como aspecto limitador a imaginação do criador do *software* destinado à representação dessa informação.

Para aclarar mais a última forma de processamento de dados (a digital), faz-se necessário um pouco mais de digressões, a fim de que os prós e contras desse ambiente de zeros e uns seja devidamente avaliado.

No exemplo citado, já não é mais novidade a existência de uma interação entre a mente humana e o sistema dedicado ao processamento dos dados digitais que foram coletados do mundo real, a fim de propiciar a carga semântica atribuída ao promontório e para garantir uma recuperação o mais consistente e confiável possível dessa informação. O sistema sensor transforma a energia de reflexão luminosa de cada ponto[173] da superfície do promontório num sinal elétrico, que pode ser expresso numa seqüência de zeros e uns, caracterizando o nível de percepção visual humana na faixa do visível, no espectro eletromagnético.

A interação citada é uma operação híbrida, podendo haver um infinito número de combinações da forma em que o promontório "é visto" pelo sistema (pode ser especialista) e, depois, como ele "é visto" (interpretado) pelo operador humano que manipula um instrumento de coleta.

Percebe-se que se esses processos de transformação não forem bem conhecidos, muita informação pode ser perdida ou alterada até a representação final do promontório numa base de dados ou quando ele for impresso em papel. É nisto que reside a drástica mudança de paradigma apregoada por BÄHR (1996) e tacitamente incorporada aos objetivos de PRADO (2001), em que ambos garantem não existir uma teoria rigorosa sobre o comportamento da propagação de erros que se originam das duas transformações do esquema da Figura 2.1, caso estejam viciadas.

Tal teoria constituiria uma forte base conceitual para as técnicas de redes semânticas existentes. Com essa preocupação em mente, BÄHR (2000), em carta, enviou uma extensão de seu artigo de 1996, em que resgatou o poder das relações nas redes, comparando quatro delas: a RNA, a rede de Delauney (RD), a rede bayesiana (RB) e a rede semântica (RS) – aspectos gerais do artigo, em que o autor se interessa mais pelas relações do que pelas entidades que povoam uma estrutura de RC, serão vistos no próximo subitem.

De forma muito mais abrangente que os autores de trabalhos da área das engenharias (Computação, Sistemas, Cartografia), BÄHR (1996) já se apegava a conceitos de autores da

[173] Este ponto não é propriamente uma abstração geométrica *adimensional*, relacionada ao termo matemático, mas um *pixel* (V. glossário).

Lingüística e da Terminologia[174], para suprir as lacunas que os enfoques matemáticos e estatísticos alastravam no estudo da informação geográfica. É neste ponto que o conteúdo deste subitem se esgota (o nível *subsimbólico* da IG), alçando-se, novamente, ao simbólico.

A causa da ambigüidade da informação geográfica, seguramente, está na transformação que se processa do nível iconográfico para o simbólico da Figura 2.1.

No nível simbólico de descrição dos objetos do mundo real, é possível concorrerem diversos tipos de processos, que não podem existir no nível anterior (iconográfico): comparações e composições multiformes entre mapas e imagens, efetuadas com o auxílio de aplicativos de alto nível, que podem ser codificados e executados em sistemas computacionais digitais. Tais aplicativos podem ser capazes de manter uma interação cognitiva com o operador humano e é sobre este rebento da IA, o fenômeno da similaridade semântica (SS) que o próximo subitem tratará.

3.2.2.2.2. Tópicos sobre similaridade semântica

Para iniciar esta parte da revisão, que é a mais relacionada com a fundamentação teórico-empírica para o problema de pesquisa deste trabalho exploratório, é conveniente citar alguns autores da Filosofia da Linguagem e até da Semiótica, porque, como se verá dos autores que investigam o fenômeno da SS pelo enfoque empírico, muitas das brechas metodológicas podem ser cimentadas por *constructos* que surgiram de reflexões desses autores-pensadores. Por conseguinte, este subitem adotará a técnica *top-down* de revisão, analisando os mais reflexivos e passando, gradualmente, aos mais formais, procurando manter a difícil concatenação racional entre os elementos desses mundos tão afastados um do outro.

E se o assunto começa pelo mais abstrato e geral (*top*), é melhor iniciar por uma breve reflexão de SANTAELLA (1983), com relação ao enfoque da Semiótica sobre informação. A autora não classifica a Semiótica como uma ciência. No início de sua obra, a autora identificou esse campo do saber como um exercício de filosofia científica da linguagem, trazendo para o seu rol de preocupações de estudo e análise o universo multiforme dos fenômenos de todas as formas de linguagem. A autora chegou a fazer uma relação desse rol de preocupações da Semiótica com a própria *vida*, numa interessante formulação de sentido figurado: **VIDA = ENERGIA ⊕ INFORMAÇÃO** {vida := linguagem; energia := processos dinâmicos informação := ajustes aos processos dinâmicos}.

---

[174] "A practical course in terminology processing", de Juan C. Sager (1990).

Apesar de ser esse caráter holístico da Semiótica a fonte de reservas e de algumas das críticas que recebe de outros pensadores, essa formulação pode ser analisada por outra perspectiva (um pouco mais objetiva), numa formulação semelhante proposta por BÄHR (2000), em que a "parcela" chamada de informação na fórmula anterior passa a ser considerada como conhecimento na deste autor.

BÄHR (2000) frisou que um termo isolado não sustenta a realidade, sendo isso válido tanto para a linguagem natural como para a imagem. Essa verificação vinha de uma reflexão do autor sobre uma leitura de um artigo de Makato Nagao[175], que a ilustrava pela relação: **CONHECIMENTO = COGNIÇÃO ⊕ LÓGICA**.

O *conhecimento* (fatos + regras) da fórmula deve começar por um conjunto de axiomas básicos. As condições de ampliar esse conhecimento básico devem residir num mecanismo de inferência (*cognição*) que teste os estímulos do MR em relação aos axiomas contidos na BC, num processo contínuo e estruturado pela *lógica*. É a lógica que permite ajustar o processo cognitivo segundo um padrão de configuração, já que há uma infinidade de instâncias cognitivas dos indivíduos humanos diante da realidade espacial.

Em seu segundo trabalho, BÄHR (2000) procurou comparar diferentes modelos de redes de representação do conhecimento, tendo como referencial o fato de que o homem usa as estruturas de dados contidas na sua mais atual ferramenta de trabalho intelectual – o computador –, para transferir o seu entendimento ou os modelos mentais do conhecimento adquirido.

Nesse ponto, uma transferência da leitura para o item "dado, informação, conhecimento e sabedoria" do glossário é bem conveniente, para analisar as visões da Engenharia de *Software* (PRESSMAN, 1995), da Computação Paralela (HWANG, 1985), da Análise e Projeto de SGBDs (SETZER, 2001) e das Ciências Sociais (URDANETA, 1992). As visões de cientistas da computação ajudam a entender o problema da semântica que se deseja incorporar a um SIG. No caso de SETZER[176] (2001), surpreendentemente, a investigação entra em considerações filosóficas sobre o assunto, o que permite entender por que o foco excessivo nos aspectos práticos às vezes limita a compreensão do problema.

As formulações figurativas de SANTAELLA (1983) e de BÄHR (2000), vistas atrás, podem ser consideradas como um marco[177] para ser atingido pelas iniciativas de pesquisa para criar um SIG de natureza semântica, cujos aspectos de ordem teórica PRADO (2000)

[175] "Knowledge and interference" (1990).
[176] A posição deste autor é francamente contrária à existência da IA, em que pese ser um cientista da computação.

conseguiu sintetizar e pôr em termos mais práticos para levantar os requisitos necessários para a construção de interfaces amigáveis para uma ampla gama de usuários de *geotecnologias*.

Voltando à análise da Figura 2.1, pôde-se inferir que as implicações lingüísticas que tomam parte na transformação do nível iconográfico para o simbólico são causadas por inter-relações de ordem lógica e topológica. O que se deseja nessa transformação é o mínimo possível de perda de informação, já que do mundo real para o iconográfico há um processo de transformação (projetiva) que não deixa de estar afetado por ruído (imprecisão dos aparelhos sensores e inadequação dos algoritmos de transformação).

A problemática lingüística no nível simbólico toma forma preocupante, se os métodos (gramáticas, grafos conceituais, redes semânticas) não levarem em conta o ***contexto***. Ainda que tais línguas sejam afetadas por um grau de ambigüidade, é pelo *contexto* que essas línguas naturais tornam-se agentes portadores de informação. Por outro lado, é também no nível simbólico que o *contexto* permite a representação inequívoca do objeto do nível icônico numa forma totalmente assimilável (compreensível) para a mente humana.

Nesse aspecto, MALMBERG (1976) analisou a trajetória paralela do desenvolvimento lingüístico e do desenvolvimento intelectual na busca de conhecimento, ao contrapor à crítica de que a língua engessa[178] o pensamento o argumento de que sua riqueza vem justamente daí, porque carrega a capacidade de abstrair e generalizar: "Da pluralidade cria unidades totais".

Ainda dentro dessa reflexão, MALMBERG (1976) foi endossado por BÄHR (2000), quando acentuou o valor do *contexto* na linguagem, ao declarar: *"Toda palavra adquire a plenitude de seu significado unicamente dentro de um contexto concreto, i.e., dentro de um conjunto de conotações e associações (texto ou enunciado) que, juntas, criam um significado. Um dicionário (glossário) não pode dar conta de todas as possibilidades semânticas da palavra: é uma lista de abstrações ... e, dentro deste contexto, vem o espectro amplo das línguas e os seus respectivos sistemas sinalativos, construídos ao longo do tempo, sobre aspectos emocionais, culturais e convencionais..."*

BÄHR (2000) pretendeu, justamente, transpor para ferramentas externas ao domínio mental humano a semântica que as representações mentais estabelecem entre os fenômenos espaciais e transferiu mais ênfase às relações do que às entidades em si. Essas ferramentas seriam quatro: RNAs, as redes de Delaunay (RDs), as redes bayesianas (RBs) e as

---

[177] Conjunto de requisitos.
[178] Porque limita a designação dos fenômenos que são muito dinâmicos no MR.

redes semânticas (RSs), como já anunciadas anteriormente. Cada uma privilegia um aspecto do relacionamento entre as entidades espaciais, que é materializado pelos vínculos (arcos) que ligam os objetos (nós) da rede. Essa *trama de arcos* constitui, segundo BÄHR (2000), *informação contextual*, que pode ser de natureza geométrica (RD), ou semântica (RS), ou estocástica (RB), ou de aprendizagem por treinamento (RNA). A escolha de qual representação adotar dependerá da aplicação em curso.

BÄHR (2000) intentou mudar o foco dessas estruturas de RC para as relações entre os nós, até então negligenciadas, segundo o autor. Os nós de uma estrutura conceitual em rede agregam elementos determinísticos, primitivas geométricas, objetos, coordenadas e termos; as ligações agregam relações de segmentação, de instância e de especialização (RSs), assim como poderiam usufruir de propriedades estocásticas (RBs).

Um termo isolado não se sustenta na realidade empírica e isto é válido tanto para expressões em LN como para imagens e mapas (BÄHR, 2000).

Assim como SÖRGEL (1999) reclamava da falta de sinergia entre profissionais da Ciência da Computação e os da Terminologia, BÄHR (2000) reclamou da mesma falta de sinergia entre estudiosos do campo do *Geoprocessamento* (imagens) e os da Lingüística.

BÄHR (2000) enumerou alguns parâmetros para orientar a escolha de um dos modelos de rede citados, partindo do princípio que o homem, ao produzir conhecimento, tenta transferi-lo, carregando os modelos que o estruturam em programas de computador (sua ferramenta). Há duas soluções para esse problema de modelagem e de transferência do conhecimento: as *implícitas* (por sistemas baseados na heurística) e as *explícitas* (por sistemas aplicados a bases de conhecimentos).

No caso das soluções explícitas, os processos de aquisição e de processamento do conhecimento são separados. No caso das implícitas, o sistema é treinado por um especialista (homem), tornando-se capaz de "aprender". Os sistemas explícitos são mais completos e dependentes da Lógica (modelos formulados *a priori*) que os implícitos.

Depois de explanar essas características gerais dos sistemas de manipulação de conhecimento, já é possível listar alguns parâmetros de apoio à escolha do tipo de rede.

Para *redes neurais*:

- Típicos sistemas implícitos;
- Princípio básico: a um dado estímulo (conceito de entrada), correspondem estímulos de saída (reações ou conceitos de saída) por meio de uma cadeia ponderada de ligações, a qual depende, num determinado momento (instância da rede), dos dados (exemplo de treinamento que vai "adestrar" as ligações) fornecidos pelo usuário humano;

Para *redes de Delauney*:

- Típicos sistemas explícitos;
- Os fenômenos modelados são comparados com malhas de triângulos (modelos *a priori*);
- Apesar de as arestas dos triângulos não constituírem relações portadoras de informação, a geometria do triângulo incorpora significado de uma forma *gestáltica*, i.e., integrada aos significados de cada geometria de um triângulo vizinho na malha.

Para *redes bayesianas*:

- Típicos sistemas explícitos;
- Os objetos da rede e suas relações são previamente estimados por probabilidade.

Para *redes semânticas*:

- Típicos sistemas explícitos;
- Os nós são conceitos interligados por relações diversas, sendo as mais comuns as hierárquicas.

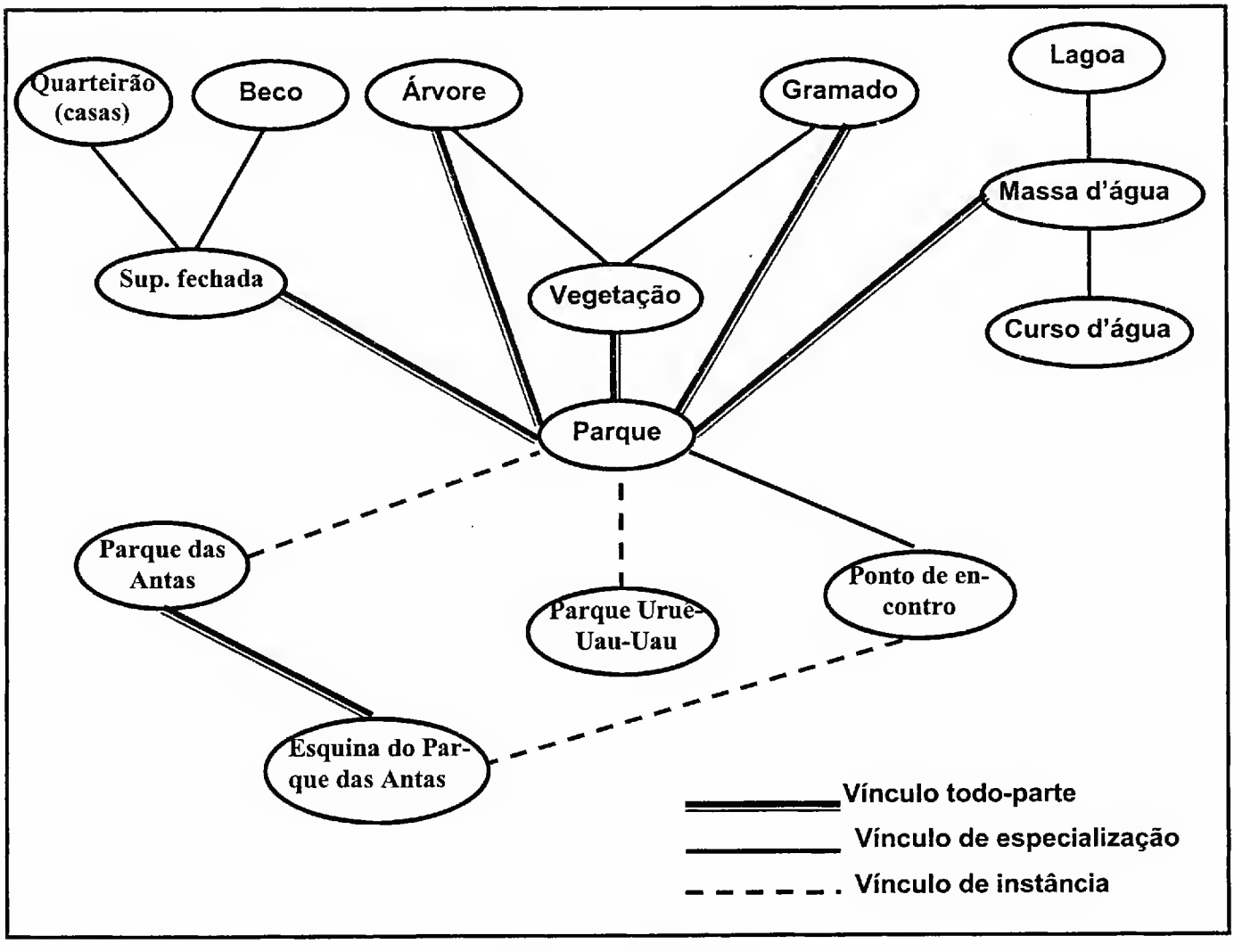

Figura 3.9: Rede semântica para objetos espaciais.

BÄHR (2000) citou F. Quint, que, em 1997, foi o primeiro pesquisador a utilizar uma RS como ferramenta de interpretação de imagem. Sua idéia central foi a de modelar uma imagem por uma estrutura portadora de significado como a RS, apesar de a imagem estar num nível muito complexo de abstração, próxima ao MR (nível iconográfico).

BÄHR (2000) frisou o poder contextual das ligações desses quatro tipos de redes, tanto para signos verbais como para signos não-verbais (multidimensionais), como as imagens.

A Figura 3.9 mostra uma ***rede semântica*** e como ela é utilizada para descrever a modelagem de partes de uma imagem (nível iconográfico) no nível simbólico. O esquema é composto de *nós* e *vínculos* (ligações ou relações). Objetos espaciais contidos numa imagem ou num mapa são representados pelos nós (termos); os relacionamentos entre esses objetos são representados por três estilos de linhas (vínculos): *relações todo-parte*; *relações de especialização* e *relações de instância* (ocorrência singular, única), pela terminologia de BÄHR (2000), facilmente traduzida para a OO, para a Terminologia científica e para a Lingüística Computacional, conforme já visto

O trecho em linguagem natural que corresponde à definição estruturada dada pela rede semântica da Figura 3.9 é o seguinte: "Vegetação, áreas fechadas e massas d'água são componentes de 'parque'. Os componentes estão conectados a esse termo pelas linhas todo-parte. De modo igual, árvores e gramados são encontrados num parque (linhas todo-parte). Por outro lado, árvores e gramados são tipos especiais de vegetação (linhas de especialização). O Parque Nacional Indígena Uruê-Uau-Uau é uma ocorrência local de 'parque' (linha de instância). A esquina do Parque das Antas é um setor do Parque das Antas (linha todo-parte). Esta esquina é um ponto de encontro, que faz parte do conceito de 'parque', como por exemplo: área de lazer infantil. A esquina do Parque das Antas é uma instância de 'ponto de encontro' (linha de instância)."

Comparada ao nível iconográfico[179], a estrutura da de rede semântica da Figura 3.9 (nível simbólico) é conceitualmente muito mais rica de conteúdo informativo e mais clara. É de evidência meridiana que tanto os objetos como as relações entre eles recebem um alto grau de influência dos aspectos terminológicos (termos e conceitos), estando sujeitos a uma carga de ambigüidade difícil de mensurar (BÄHR, 1996).

Quando BÄHR (1996) concluiu por isto, não sabia que a dificuldade ainda perduraria por mais cerca de oito anos (até hoje), porém com possibilidades de mensuração. Ele mesmo, no seu artigo de 2000, apresentado no Congresso Internacional de Fotogrametria de Amsterdã (ISPRS), ensaiou uma forma de avaliação de semântica da IG, propondo a com-

paração de um aspecto comum aos quatro tipos de rede já enumerados, até então um tanto negligenciado[180] pelos engenheiros e cientistas das exatas, qual seja: *as relações entre conceitos*, mais estudadas na Lingüística e na Psicologia Cognitiva.

Essa comparação é importante para delimitar o mecanismo de representação do conhecimento que se vai utilizar nesta pesquisa: a RS, na forma de uma estrutura de árvore, na qual se pode admitir que houve inserção de informação contextual, tanto no primeiro modelo, implementado pelo PROFAX, em que o contexto, de certa forma, está embutido nas ligações entre as entidades e em que tais ligações foram modeladas pela função matemática do co-seno entre dois vetores não-nulos no espaço vetorial euclidiano, assim como no segundo modelo, implementado pelo PRONTO©, em que a ponderação idêntica atribuída a cada um dos três coeficientes da fórmula geral de similaridade semântica de RODRÍGUEZ (2000) é coerente com o campo de observações que envolve o espaço de estimulação sensorial dos indivíduos respondentes.

A função do co-seno (GANESAN, 2001), no primeiro protótipo, e as fórmulas[181] de RODRÍGUEZ (2000), no segundo protótipo, representam o fenômeno da SS; e é precisamente sobre essa SS que a revisão a seguir vai tratar, contudo, de forma muito abreviada, apenas para dar o necessário suporte aos trechos da seção de metodologia que virão no Capítulo 6.

Das quatro estruturas analisadas por BÄHR (2000), tanto MEDEIROS (1999) como RODRÍGUEZ (2000) encontraram nas RSs a forma mais adequada de organizar os seus sistemas conceituais de base lingüística. Esta pesquisa também adotou essa estrutura de RC, denominando-a de *árvore n-ária* numa etapa de maior proximidade da fase de implementação de um dos instrumentos de pesquisa – o PRONTO©.

Segundo MEDEIROS (1999), as redes semânticas (RSs), de largo emprego na IA, foram propostas por M. R. Quillian, em 1968, como uma forma de representar significados de palavras inglesas que denotassem tanto objetos como eventos.

Nas redes semânticas, os conceitos são representados como *nós* e as relações entre eles são representadas por *arcos*. Cada arco possui uma etiqueta associada, que determina a relação entre os nós conectados.

As relações estabelecidas com maior freqüência são as de gênero-espécie ("x" é um tipo de "y") e as de todo-parte ("x" é parte de "y").

---

[179] Imagine uma fotografia desse parque.
[180] Novamente vem à tona a observação de SÖRGEL (1999) sobre a falta de sinergia nesses campos do conhecimento.
[181] Função do co-seno na Eq. 3.1 e fórmulas de RODRÍGUEZ (2000) da Eq. 6.1à 6.3, no Cap.6.

Outras relações podem ser inseridas na rede semântica. Os tipos de relações que se encontram na literatura são bastante variados e dependem do domínio da aplicação.

As redes semânticas baseiam-se na noção de atributo-valor, que são um padrão em muitas aplicações da Ciência da Computação. O nome do atributo é visto como o nome do elo ou a referência da rede semântica que faz a ligação com o valor do atributo.

Segundo G. Sabah [*apud* MEDEIROS (1999)], a facilidade de realizar deduções por meio de composição de relações é uma das razões pela qual as redes semânticas se tornaram tão populares. A herança de propriedades evita a repetição da representação dessas propriedades a todos os membros de uma cadeia hierárquica do tipo gênero-espécie.

Por outro lado, a herança de propriedades pode causar falsas interpretações quando determinados indivíduos não herdam todas as características de sua classe. Nestes casos são necessários conhecimentos suplementares para resolver eventuais contradições.

Nas redes semânticas, as características dos conceitos são informadas por meio de traços semânticos (características e feições distintivas - *fds*).

Como ressalta Lepage [*apud* MEDEIROS (1999, p. 93)], os traços semânticos funcionam como critérios semânticos que validam a aceitabilidade de um enunciado. É pela análise desses traços que os enunciados *"A mesa ensina"* e *"Comi um tijolo"* são rejeitados semanticamente. As ações de ensinar e de comer exigem, respectivamente, um sujeito com o traço +humano e um objeto com o traço +comestível.

No trabalho de MEDEIROS (1999), os conceitos encontram-se estruturados em rede semântica, segundo uma ontologia proposta para o Mercosul. No trabalho de RODRÍGUEZ (2000), a estrutura de organização dos conceitos é semelhante, mas a fonte terminológica para montar a ontologia foi extraída de duas linguagens documentárias: o catálogo SDTS™ (*Spatial Data Transfer Standard*) e a taxinomia *Wordnet*™, que se complementam (V. subitem 6.2).

O diagrama de conceitos definido por FELBER (1984) pode assumir as mais variadas denominações noutros campos do conhecimento que não sejam a Terminologia, mas não diferem essencialmente na sua finalidade, que é a de estruturar uma rede conceitual por meio de termos e das relações entre esses termos.

MEDEIROS (1999) utilizou a técnica dos grafos conceituais (GCs), desenvolvida por John Sowa, em 1968, quando escreveu um trabalho de final de curso para Marvin L. Minsky. Neste trabalho, Sowa aplicou a idéia de fluxogramas para criar um modelo de representação de conhecimento em IA, baseando-se em figuras de caixas e círculos para gerar gráficos conceituais. Na década de 70, Sowa iniciou um trabalho sério de pesquisa sobre gráficos

conceituais como linguagem de representação do conhecimento no *Systems Research Institute*, da IBM, culminando, em 1984, na Teoria dos GCs, publicada no seu livro *Conceptual Structures.*

Como observa G. Sabah [*apud* MEDEIROS (1999)], apesar das inúmeras teorias e fontes de informação citadas por Sowa em seu trabalho, a origem de suas idéias encontra-se na Psicologia da Visão, mais especificamente no conceito de percepção, contribuindo bastante com a noção de conhecimento na IA: *"... é a habilidade de formar um modelo mental que represente adequadamente as coisas, assim como as ações que podem ser por elas ou sobre elas executadas".*

Como se viu até este ponto, as tentativas de vários pesquisadores e cientistas que atuaram na busca de uma forma de valorizar o significado "transportado" pelas expressões lingüísticas, sempre apontaram para uma Semântica de uma entidade do MR bem definida: estudo do significado por intermédio da análise dessas expressões quanto à sua denotação e *intensão.*

Recordanso PINTO (1977), os significados são categorias de entidades portadoras de informação, interpessoais e sociais, que garantem o funcionamento da língua como instrumento de comunicação. Esta definição integra as várias visões de filósofos da linguagem e semanticistas sobre o conceito de *significado*.

O autor citado, adepto convicto da linha de pesquisa caracterizada pela "quantificação" do significado, preconizou uma fórmula para a sua avaliação (relativa), chamando-a de ***cálculo*** de ***intensões*** ou de ***predicados***, por ser a *intensão* ou conteúdo informativo a fonte mais viável e plausível de investigação para a Lógica.

PINTO (1977) enumerou dois objetivos para o cálculo de *intensões* (de predicados, análise semântica ou análise *componencial*):

- De cunho semântico: precisar a noção de conotação ou *intensão*;
- De cunho sintático: permitir o cálculo da intensão de um signo complexo, sintagma ou enunciado, com base na *intensão* de seus componentes.

O campo da Semântica interessado nesse cálculo se chama de Semântica Estrutural (PINTO, 1977, p.70), cujas origens vêm de L. Hjelmslev, por intermédio de A. J. Greimas.

A Semântica Estrutural está interessada, portanto, na análise semântica de vocábulos da língua, organizados por campos semânticos, por meio de traços ou feições distintivas pertinentes.

Trata-se de um estudo que decompõe o significado de um conjunto de expressões (termos), que se supõe terem algum parentesco semântico nas suas características definido-

ras; cada vocábulo deve ser analisado quanto às diferenças e semelhanças de significado que o relacionam com os demais e que sejam pertinentes para um falante e seu interlocutor numa atividade de comunicação. É, por conseguinte, um cálculo de *intensões* que obedece aos três postulados já vistos (p. 145) do realismo crítico da Filosofia da Linguagem e que já vinha sendo praticado desde meados da década de 50 do séc. XX, segundo PINTO (1977), tendo certa tradição na Antropologia Social, quando o escopo do estudo era restrito às relações de parentesco entre grupos humanos. O mesmo autor, no entanto, regride ainda mais no passado e localiza na obra de Jost Trier, editada em 1931, a primeira iniciativa de cálculo de *intensões*.

PINTO (1977) considerou a metodologia de solução de problemas de análise sêmica de campos semânticos muito difundida, mas discutiu alguns pontos de ordem teórica, muito apropriados para receberem a devida atenção neste trabalho.

Na verdade, o que o autor acima chamou de *campo semântico* é o tema da LN escolhido como *corpus* do experimento e que estará sujeito a um processo de classificação tradicional como o da *análise por facetas* (VICKERY, 1980). Logo a seguir, entra em jogo um processo de construção de ontologias para cada classe de entidades de mais alto nível, em que a definição por *intensões* é o cerne deste processo. Destarte, a noção de análise de campos semânticos de PINTO (1977) é semelhante à análise por facetas de VICKERY (1980).

Em geral, para estruturar as categorias que se formam dentro do campo semântico em trabalho, monta-se uma RS ou uma estrutura taxinômica (árvore), na qual os nós-folhas são os significados mais específicos sobre os quais se deseja fazer o cálculo de predicados ou análise sêmica. Não quer dizer, com isso, que não se possa realizar tal análise em níveis mais altos da estrutura, mas como o interesse é no cálculo das *intensões* e como estas estão explicitadas nos níveis mais baixos da taxinomia, é nesses níveis que o cálculo será determinante.

Exemplificando (PINTO, 1977, p.72), sejam as três categorias seguintes, na ordem decrescente de generalização: ENTIDADE CONCRETA – MATERIAL – VEGETAL. Isto representa um *campo semântico*. Para RODRÍGUEZ (2000), isto é um *diretório semântico*. Para VICKERY (1980), os três termos são as *categorias fundamentais*. PINTO (1977) identificou quatro significados para este campo semântico: *madeira*, *árvore*, *bosque* e *floresta*. Cada significado é composto de *características definidoras*, *traços semânticos*, *funções distinti-*

vas[182], *semas* ou *marcadores sintáticos*, dependendo do autor. Ao conjunto de semas de cada significado, PINTO (1977) denominou de *semema*.

Semas e sememas são classes de sistemas de categorização da LN a que pertence o falante. Os sistemas formados por semas são denominados de *sistemas de categorização primários* e os de sememas, de *secundários*. O *contexto* de semas e sememas é o *campo semântico* que o falante experimenta[183] (presta a atenção).

Outra entidade lógico-matemática que surge das combinações de semas para formar sememas é o chamado *produto lógico* (PINTO, 1977).

O problema fundamental da Semântica Estrutural está relacionado às montagens dos sistemas primário e secundário, podendo-se pô-lo na forma da seguinte pergunta: "O grupo de usuários ao qual se destina a taxinomia vai conseguir assimilar, sem óbices, o sistema de significados imanente ao campo semântico determinado pelo pesquisador?"

Como disse PINTO (1977), é claro que existem outros sistemas de categorização primário e secundário, capazes de organizar o mesmo material, logo, por que privilegiar um sistema no lugar de outro?

Para amenizar essa divergência de enfoques na opção por um sistema de categorização, PINTO (1977) chamou a atenção para as mesmas regras básicas enumeradas por VICKERY (1980), RICH (1993), RUSSELL (1995), FURLAN (1998) e muitos outros autores. Essas regras básicas são:

- Classes mais genéricas têm mais oportunidade de fazer parte do domínio cognitivo do falante do que as classes específicas.
- O sema deve ser escolhido de forma a permitir ao falante fazer as conexões semânticas adequadas entre as expressões ou termos que com ele compõem o semema, ou seja, ao escolher um sema como "beber", é apropriado que o outro sema que participe do sistema primário de categorização seja algo com as características de "líquido".

Na regra anterior, cabe uma explicação: a escolha de semas baseados nesses critérios de conformação semântica, entre os falantes que observam um campo semântico de sua LN ou língua de especialidade, é chamada de "*restrição de seleção*" por PINTO (1977).

- Para tornar mais fácil o processo de restrição semântica na escolha de semas, é aconselhável propor um conjunto de testes preliminares aos indivíduos que serão submetidos à

[182] Segundo RODRÍGUEZ (2000)

[183] Há 3 níveis de contacto do pesquisador com a realidade: experiência (atenção), observação (atenção + rigorosa com coleta de dados) e experimento (nível mais alto, em que o pesquisador interfere na relaidade).

avaliação de um campo semântico, a fim de garantir a adequação empírica da escolha dos semas.

PINTO (1977) citou autores como B. Bierwisch e A. J. Greimas, que postularam a existência de "universais semânticos", de uma maneira próxima ao kantismo. Esses *universais semânticos* seriam capazes de descrever as relações semânticas entre termos e expressões de uma língua e estariam situados num nível estável (invariante) da faculdade cognitiva e perceptiva do ser humano. Nesse caso, o *constructo* sema seria o mais adequado para representar os *universais*, mas PINTO (1977) é céptico com relação a essas tendências universais de sistematizar a teoria semântica nesses primórdios de investigação formal.

Nesse ponto de desenvolvimento das pesquisas, PINTO (1977) estimou que os semas são expressões da língua, tomadas momentaneamente, a fim de determinar o valor do significado de outras expressões. É uma visão etnocentrista[184] do conceito de sema.

Mas o maior problema que um pesquisador pode enfrentar na tarefa de estimar o valor do significado de uma expressão lingüística é o de delimitar o *campo semântico* em que se insere a expressão. A *delimitação de um campo semântico* se inicia pela aquisição do conhecimento intuitivo do significado de um vocábulo (PINTO, 1977).

Há muita preocupação de cientistas em relação à base teórica para a delimitação do campo semântico. T. Todorov [*apud* PINTO (1977, p. 76)], p.ex., foi pessimista quanto ao estabelecimento de uma metodologia segura para fixar os limites de campos semânticos, mas não ofereceu alternativas. E. Coseriu [*apud* PINTO (1977, p. 76)], p.ex., sugere que essa delimitação se faça por meio do conceito de *paradigma*: "Farão parte do mesmo campo semântico todas as expressões que podem ser comutadas em algum ponto da cadeia sintagmática". Esta posição se assemelha à Teoria do Protótipo de Eleanor Rosch (V. Quadro 3.2 e Figura 3.11).

Com respeito à posição acima, PINTO (1977) afirmou ser uma alternativa inviável em matéria de formalização, desde que, para formar paradigmas (padrões), somente seriam derivados elementos paradigmáticos de posições do sintagma[185] que fossem marcadas semanticamente pela intenção do usuário, com uma carga alta de subjetividade, o que seria uma fonte de vício para o processo de cálculo.

A posição mais sóbria e objetiva, segundo PINTO (1977), é a de B. Pottier e de E. H. Bendix. Estes autores entendiam que a delimitação depende do ponto de vista do pesquisador do campo da Lingüística, fundando-se principalmente em suas intuições. Os testes de

---

[184] Tendência de admitir a cultura de um povo como referencial de estudos sociais aplicados.
[185] Unidade lingüística dotada de significado. Os sintagmas são elementos construtores de frases.

adequação empírica e a análise dos resultados sancionarão ou não a opção metodológica adotada com base na intuição do usuário, o que pode ser uma fonte de vício no processo, não o tornando imune a essa inevitável tendência. Essa opção foi também compartilhada por RODRÍGUEZ (2000).

Outra posição que complementa a anterior é a compartilhada por A. J. Greimas e M. Bunge. Ambos consideraram primordial um trabalho de convergência heurística sobre o campo semântico, i.e., um contexto o mais fechado possível em termos formais e semânticos, para que seja possível efetuar os cálculos de *intensões* (ou, por extensão, de avaliação de SS). Esta, posição volta-se a frisar, é também compartilhada por autores da IA, tais como: RICH (1993), RUSSELL (1995) e MEDEIROS (1999).

A validação intuitiva dos indivíduos que serão avaliados sobre o campo semântico escolhido pelo pesquisador é um fator primordial para a delimitação desse campo.

Na presente pesquisa, foi este mandamento que determinou a seleção de indivíduos e do *corpus* que contém o campo semântico[186] que se pretendeu estudar.

PINTO (1977) continuou a enumerar algumas recomendações sobre a delimitação de campos semânticos de *corpora* em texto-livre (LN), o que não se aplica ao estudo-de-caso em questão, que trata de uma língua profissional, muito mais controlada do que a LN em matéria de ambigüidade.

A análise de *intensões* que o pesquisador deverá levar a cabo basear-se-á no cálculo de semelhanças (similaridades) e diferenças avaliadas no interior de um campo semântico. É crucial, para esse cálculo, delimitar a quantidade de expressões que o compõem, visto que, do contrário, há o risco de se cair numa regressão infinita (PINTO, 1977) ou num problema NP (RUSSELL, 1995).

Daqui até o final deste subitem, serão levantados apenas os principais tópicos da revisão de literatura ligados às teorias de cognitivas que mais contribuíram para a avaliação da SS e para o embasamento teórico que serviu para consolidação dos objetivos específicos de pesquisa e das hipóteses estatísticas, formulados a seguir. Desta revisão, em ordem cronológica, serão estudados (isoladamente ou em grupo): RIPS (1973), TVERSKY (1977), KRUMHANSL (1978), RADA (1989), MILLER (1991), JIANG (1997), RESNICK (1999), WONG (2000), GANESAN (2002) e SANTOS (2002).

É bom frisar que RODRÍGUEZ (2001) não foi citada por ser considerada a autora de consolidação, porque, exceto pelos três últimos autores acima, de todos os outros ela retirou

---

[186] Folha Faxinal© e a documentação do seu modelo conceitual.

fundamentação teórica direta para a sua tese. Os trabalhos mais antigos (décadas de 70 e 80) foram enviados por RODRÍGUEZ (2001) por malote postal a este pesquisador, a fim de cobrir algum ponto pouco desenvolvido em sua tese ou até mesmo para despertar interesse de pesquisa em assunto por ela descartado.

Nessa condição, como esta tese tem sua origem na de RODRÍGUEZ (2001), logo a seguir vêm relacionados os pontos mais relevantes do trabalho desta autora:

- Problema de pesquisa (em 5 questões): "Quais são as propriedades que deve possuir um modelo de similaridade para classes de entidades espaciais?"; "Quais são as características de definição de classes de entidades espaciais?"; Quais são as vantagens e desvantagens dos modelos correntes de avaliação de similaridade semântica (SS)? As vantagens desses modelos correntes podem ser integradas num novo modelo? Como o contexto influi na avaliação da SS? (RODRÍGUEZ, 2001, p.13 e 14);
- Objetivo geral: "Criar um modelo de avaliação da similaridade semântica entre classes de objetos espaciais, que reflita as propriedades do senso humano de julgamento de similaridade e que seja implementado por um sólido formalismo computacional";
- Pressuposto de pesquisa: "O modelo de similaridade semântica (MSS) alcança o senso humano de julgamento dessa similaridade" (RODRÍGUEZ, 2001, p.14);
- Escopo e fundamentos teóricos: "Este trabalho tem seu foco sobre as entidades espaciais expressas por termos em vez de relações espaciais expressas por proposições em língua natural... A modelagem matemática originou-se nos estudos de Psicologia Cognitiva e do PLN... A avaliação de SS é um processo que analisa feições comuns e distintas entre classes de entidades, além das relações de generalização e de composição entre estas entidades (relações semânticas). Um sistema conceitual assim construído constitui uma ontologia".
- *Corpus* da tese: montou sua ontologia com base em duas linguagens documentárias: o catálogo SDTS™ (*Spatial Data Transfer Standard*) e a taxinomia *Wordnet*™, que se complementam (V. subitem 6.2).

O trabalho de RODRÍGUEZ (2000) foi uma alternativa híbrida entre os quatro tipos gerais de modelos de avaliação de similaridade semântica. A própria autora estabeleceu uma tipologia para esses grupos. De maneira geral, os 4 modelos são os seguintes:

(1) **Modelo de feições:** originado nos estudos de TVERSKY (1977). Associa boa parte da Teoria dos Conjuntos aos resultados obtidos em testes empíricos elaborados por psicólogos cognitivos. Esse modelo tem a função de complementar o poder discriminatório do modelo que vem a seguir, ao estabelecer mais pormenores de propriedades e funcionali-

dades para cada classe de entidades. Esses pormenores são chamados ***de feições distintivas (fds)*** e abrangem: ***partes***, ***funções*** e ***atributos***. Pela visão lexical, partes são substantivos, funções são verbos e atributos, adjetivos;

(2) **Modelo de relações semânticas:** baseado em teorias da Ciência da Computação e tem nas RSs o seu instrumento básico de modelagem. Esse modelo contempla as relações de gênero-espécie e de composição entre as entidades, estabelecendo uma ordem hierárquica entre suas classes de representação;

(3) **Modelo de conteúdo informativo:** baseia-se na Teoria da Informação, em que o conteúdo informativo[187] de uma determinada classe de entidades é inversamente proporcional à taxa de freqüência das instâncias dessa classe num dado domínio de ocorrência;

(4) **Modelo baseado em contexto:** vem das teorias lingüísticas, especialmente do trabalho de MILLER (1991). São conceitos muito complexos, porque há uma carga de subjetividade muito grande nessa forma de avaliar o fenômeno da SS. Basicamente essas teorias se expressam da seguinte forma: "Para palavras do mesmo idioma, retiradas de categorias de mesma ordem sintática e semântica, quanto mais duas dessas palavras possam ser substituídas dentro de um mesmo contexto, mais similares em significado elas serão".

RODRÍGUEZ (2001) derivou seu MSS dos modelos (1), (2) e (4). Do (1) para o (4), decresce a intensidade de influência do modelo no MSS. Nessas condições, a autora até especifica um público-alvo para os resultados de seu trabalho. Basicamente, pesquisas em SRIs e em discriminação de objetos espaciais seriam áreas voltadas ao seu trabalho. A autora também incluiu nesse público de interesse grupos multidisciplinares de pesquisa em Geografia e Ciência da Computação. Mais especificamente, a autora listou projetistas de sistemas de bancos de dados e de linguagens de busca espacial, pesquisadores de IA, de sistemas de informação com ênfase em interoperabilidade, de PLN e cientistas cognitivos.

O PROFAX utilizou predominantemente o modelo (2), justificando-se em MILLER (1991) e BÄHR (2000) alguma contribuição de contexto do modelo (4) para o protótipo. Na fase de acabamento do PROFAX, foi um dos intentos adicionar alguma funcionalidade do modelo (1) preparando-o para a evolução do PRONTO©.

Daqui em diante, serão vistos os resumos dos trabalhos dos dez autores citados.

RIPS (1973) e TVERSKY (1977) são autores que trabalharam exaustivamente em testes psicológicos de natureza empírica e derivaram formalismos matemáticos para os seus

---

[187] Indicador de SS.

modelos (A. Tversky apoiou a sua tese sobre ***fds*** na Teoria dos Conjuntos e na Álgebra Linear).

O trabalho de RIPS (1973) tratou basicamente de investigar as relações semânticas que se desenrolam na mente humana. Sua base está no **modelo de memória semântica** de QUILLIAN (1968), que estabeleceu um modelo conceitual para explicar os processos que envolvem a capacidade de memória humana por uma rede de nós (conceitos) e arcos (relações lógicas entre os conceitos).

A tese de M. R. Quillian foi a de que o seu modelo poderia ser implementado num sistema computacional. A idéia geral era formar uma treliça ou árvore de conceitos com base num nó-raiz ou patriarca, que se expandiria (para baixo) por meio de arcos (disjunção, inclusão, interseção, etc.), criando nós-filhos. A expansão de cada nó "dispararia" outros nós descendentes e suas respectivas relações, que constituiriam um plano de memória semântica. Cada plano corresponderia ao *conceito simples* representado pelo nó ascendente do plano, referenciado por um termo (palavra).

O interessante nesse modelo está no fato de que cada um dos nós descendentes do plano de memória não fica isolado do resto da árvore. Todos possuem ligações externas ao seu plano de memória e é a agregação de todos esses planos de memória que representa a memória semântica de Quillian.

Por esse modelo, um *conceito completo* de um termo transcende um conceito simples limitado pelo seu plano de memória. Para se determinar um conceito completo é preciso rastrear todo o modelo (árvore), reconstituindo todas as ligações possíveis, como se fosse montar um dicionário, em que a definição de um verbete fosse determinada por outro e assim sucessivamente, de uma forma sempre recursiva[188].

M. R. Quillian também criticou[189] os modelos lingüísticos de memória, especialmente os de N. Chomsky (Gramática Gerativa-Transformacional), J. J. Katz e G. Lakoff, porque não identificou um esforço de pesquisa desses autores no intuito de agregar material semântico aos seus modelos. Quillian achou que o principal equívoco desses lingüistas foi o de dissociar o *uso da linguagem* das considerações teóricas mais nucleares de seus modelos.

Voltando a RIPS (1973), sua contribuição na teoria da SS foi a de estudar o que denominou ***efeito de subconjunto*** ou da ***distância semântica*** entre conceitos. O enunciado aparentemente complicado para descrever esse efeito parece-se muito com a já enunciada

---

[188] Evidentemente, com um patamar ou limite (*threshold*), para não descambar para um problema NP (sem solução).
[189] De forma semelhante a SCHANK (1995).

Lei de Tobler (subitem 3.1.2.4): *"No mundo, todas as coisas se parecem; entretanto coisas mais próximas são mais parecidas que aquelas mais distantes".*

Não é difícil perceber a estreita ligação do trabalho de RIPS (1973) com a ***Teoria dos Protótipos*** de Eleanor Rosch.

RIPS (1973) e sua equipe de pesquisa, para entender o fenômeno da SS, criaram uma exaustiva bateria de testes do tipo estímulos-respostas, que associavam a demora ou a rapidez da resposta de um indivíduo com o grau de similaridade menor ou maior, respectivamente, entre três termos colocados em duas sentenças. Para sintetizar todos esses testes e a conclusão de RIPS (1973) pelo *efeito de subconjunto ou de distância semântica*, seja o seguinte exemplo de sentenças contendo os termos pardal, pássaro e animal:

- Um pardal é um pássaro.
- Um pardal é um animal.

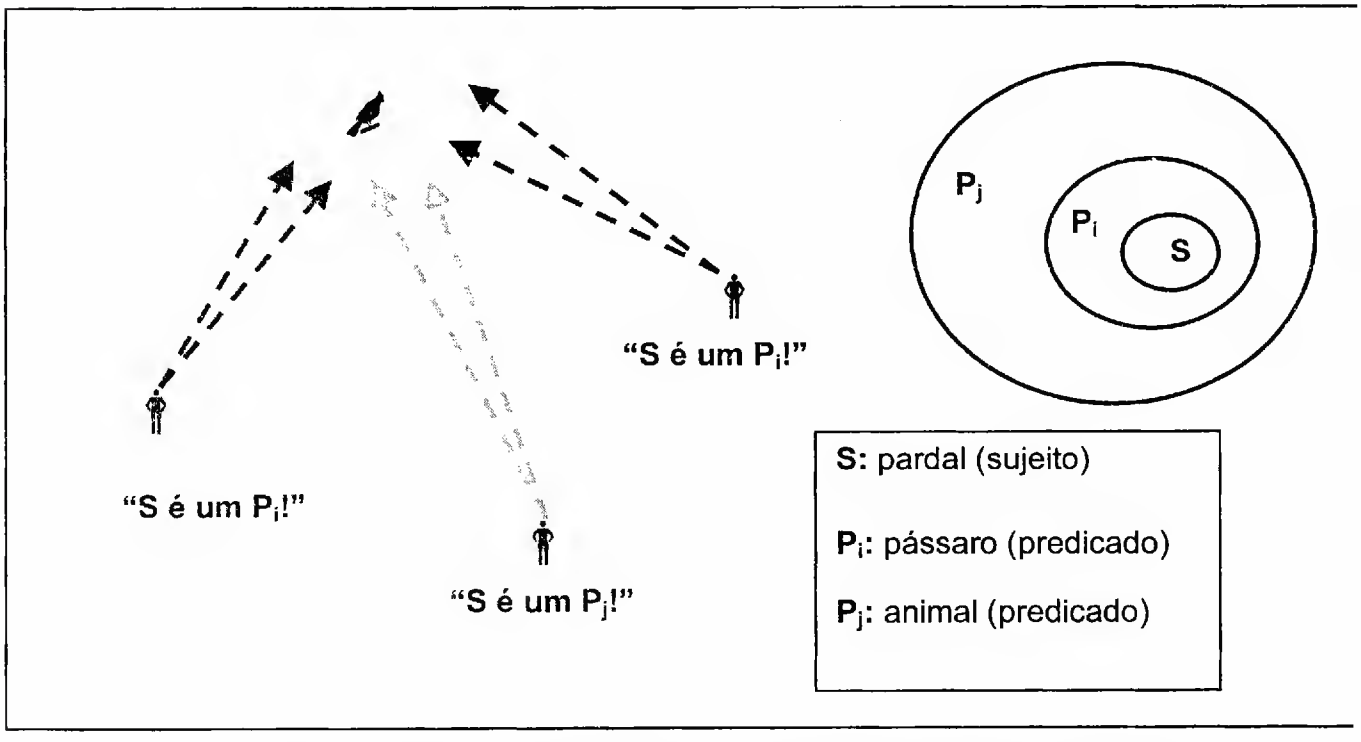

Figura 3.10: Esquema dos testes sobre o efeito da distância semântica.

A maioria das respostas foi mais rápida na identificação da primeira sentença, por ser a mais razoável.

O *efeito de subconjunto ou distância semântica* de RIPS *et al.* (1973) pode ser assim enunciado: "Quando um subconjunto é usado como predicado, as representações mnemônicas que relacionam este predicado ao correspondente sujeito são mais prováveis de ocorrer

se o predicado pertencer a um grupo mais restrito de termos (restrito em extensão) do que se o predicado pertencer a um grupo mais genérico de termos".

O esquema da Figura 3.10 ilustra melhor a idéia dos testes de RIPS (1973). Nele, dos três indivíduos inteligentes, apenas um identifica um pássaro como sendo um animal. Percebe-se que é mais fácil (freqüente) identificar um pardal como um pássaro (subconjunto mais restrito, mais próximo de pardal) do que como animal. Destarte, numa rede semântica de termos, os que estiverem mais próximos entre si na árvore representativa da rede são os mais similares semanticamente.

TVERSKY (1977) tratou de questões que envolvem dimensões espaciais na avaliação da SS. Sintetizando o trabalho desse autor, considerado como o marco teórico sobre SS, pode-se dizer que as entidades espaciais são representadas como coleções de feições distintivas (características ou traços) e a similaridade semântica se daria por uma processo de equiparação dessas feições. Sobre a SS de TVERSKY (1977)[190]:

- O conceito de SS desempenha um papel fundamental nas teorias sobre conhecimento e comportamento da Psicologia Cognitiva;
- A SS equivale a um critério de ordenação;
- Com base na SS, indivíduos podem classificar objetos e fazer generalizações;
- A SS pode ser quantificada por variáveis que estejam relacionadas com estímulos e respostas (reações) no processo de aprendizagem;
- O conceito de SS pode complementar o modelo de memória semântica de Quillian e ajudar a desenvolver uma teoria sobre reconhecimento de padrões na IA;
- A SS está normalmente associada à ordenação de pares de coisas, à busca (seleção) de coisas, ao estabelecimento de graus de semelhança entre associações, à determinação de uma escala de erros em tarefas de substituição e à correlação entre eventos;
- Similaridade não é o mesmo que a formal relação de equivalência;
- A relação de equivalência obedece a três princípios da Análise Linear. Segundo KREIDER (1972, p.46), a relação de equivalência, muito comum em diversos ramos da Matemática, atende aos seguintes princípios: 1) *Reflexividade*: $x \sim x$, para todo $x \in S$; 2) *Simetria*: $x \sim y \Rightarrow y \sim x$; 3) *Transitividade*: $x \sim y$ e $y \sim z \Rightarrow x \sim z$, sendo S um conjunto de pares ordenados (x,y) de elementos;
- TVERSKY (1973), adaptou o conceito matemático de equivalência para estudar do fenômeno da SS pelo enfoque geométrico e representou os termos por vetores num espaço euclidiano. Assim, a equivalência foi expressa pelos seguintes axiomas: 1) *Minimidade*

$\delta(a,b) \geq \delta(a,a) = 0$; 2) *Simetria*: $\delta(a,b) = \delta(b,a)$; 3) *Desigualdade triangular* ou de Cauchy-Schwarz (LIPSCHUTZ, 1972): $\delta(a,b) + \delta(b,c) \geq \delta(a,c)$

- A similaridade, por sua natureza cognitiva, não atende a todos os princípios acima, de maneira determinística, o que fez TVERSKY (1977) dar mais atenção ao *modelo de feições distintivas* (Teoria dos Conjuntos) do que ao da *distância geométrica*, cuja importância fora resgatada no trabalho de KRUMHANSL (1978);
- Vários foram os exemplos oferecidos que dissociavam a SS dos princípios e axiomas de equivalência. Com relação ao princípio da transitividade, seja o exemplo: "Embora a Jamaica seja bem similar a Cuba e Cuba seja bem similar à URSS; a Jamaica, de maneira alguma, é similar à URSS".

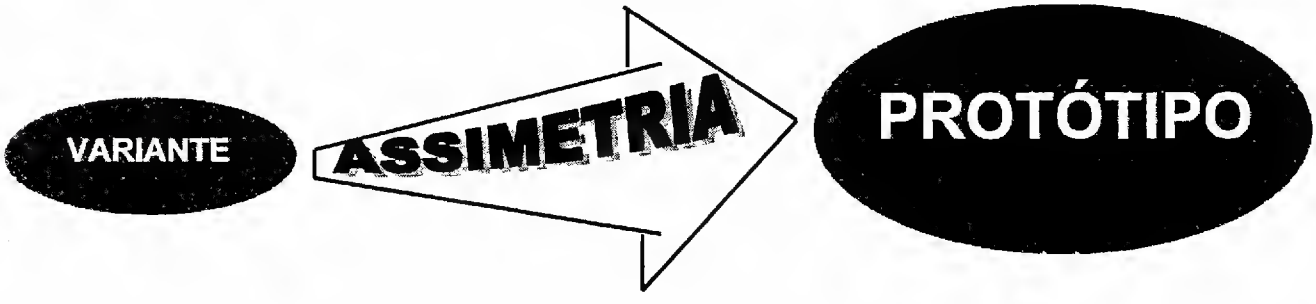

Figura 3.11: Sentido do crescimento da SS pelo efeito da assimetria.

- No caso da simetria, os testes de TVERSKY (1977) provaram um grande efeito de assimetria nas relações de similaridade entre conceitos. O termo variante (sujeito) é mais similar ao protótipo[191] (referente, predicado ou complemento) do que a recíproca. Por exemplo: 1) "O filho se parece com o pai" é uma sentença mais freqüentemente pontuada nos testes de estímulo-resposta que a sentença: "O pai se parece com o filho". Igualmente: "Um retrato representa uma pessoa" e "uma pessoa representa um retrato"; ou "Uma elipse é parecida com um círculo" e "Um círculo é parecido com uma elipse";
- Nos pares de sentenças anteriores, as primeiras estão mais carregadas de similaridade do que as segundas, mostrando um efeito crescente e direcional do variante para o protótipo, como ilustra a Figura 3.11;
- Portanto, não se deve confundir a relação de ordem cognitiva da SS com a formal de equivalência, em que pese, em discurso de senso comum, ser tolerada a confusão. Como

[190] Tópicos mesclados com comentários baseados em trabalhos mais recentes de outros autores.
[191] Pela Teoria dos Protótipos de Eleanor Rosch, o termo à esquerda da sentença é o estímulo-variável ou variante; o da direita é o protótipo ou referente. "Protótipo", aqui, não tem nada a ver com "protótipo" em Engenharia de *Software*.

diz SONESSON (2002): "O signo é assimétrico e não-reflexivo, não se podendo defini-lo em termos lógicos pela relação de equivalência. Confundir similaridade com equivalência num contexto de rigor científico, seria admitir que o homem vive no mundo abstrato das ciências exatas. No entanto, o homem vive num meio sócio-cultural bem mais complexo que o anterior, em que a SS é um fenômeno de ordem cognitiva, já observado e sujeito a muitos testes de estímulo-reação por Eleanor Rosch (1975) e Amos Tversky (1977)";

- A consideração de SONESSON (2002) não significa excluir a SS das necessárias comparações que se façam necessárias com princípios e axiomas da Matemática, como até recomendou PINTO (1977). Pelo que se depreendeu, foram apenas colocadas as devidas limitações de ordem metodológica para se analisar o fenômeno nas suas reais dimensões.

Na ordem cronológica estabelecida, KRUMHANSL (1978) levantou uma discordância com relação às restrições de TVERSKY (1977) contra os modelos geométricos[192] de avaliação de SS. De certa forma, o modelo distância-densidade de KRUMHANSL (1978) completou algumas lacunas da tese de A. Tversky sobre SS, garantindo, por evidência empírica, que o enfoque de representação da SS por modelos geométricos é adequado para determinadas situações, como as exploradas posteriormente por WONG (2000), GANESAN (2001) e SANTOS (2001), tendo sido aplicadas na implementação do PROFAX.

RADA (1989) e sua equipe provaram que é possível associar o conceito de SS aos fundamentos de uma rede semântica, dando realmente muita base para os modelos geométricos de SS: "Quanto mais propriedades em comum possuam os objetos, mais ligações e similaridade entre si eles terão numa RS". Para dar suporte experimental ao seu trabalho, RADA (1989) montou um sistema de conceitos com base em catálogos e classificações da área médica, utilizando uma metodologia semelhante à da construção de um *tesauro*.

Nesta revisão de literatura, além das conclusões de KRUMHANSL (1978), que também defendeu o enfoque geométrico para a SS, RADA (1989) foi o mais veemente defensor desse enfoque, ao admitir que a SS pode ser determinada por métricas da Geometria Analítica, levando-se em conta o conceito de pontos (vetores) num espaço *n-dimensional* euclidiano, ligados por relações hierárquicas de generalização e outras.

O autor em pauta admitiu que as exceções colocadas por A. Tversky para os axiomas da simetria e da desigualdade triangular são mais fruto da incorreta compreensão da essência da SS num contexto vetorial, porque, no exemplo dado sobre comparação entre Cuba,

---

[192] O que foi utilizado no desenvolvimento do PROFAX.

Jamaica e URSS, não foram levadas em conta certos aspectos (geográficos e políticos) nas sentenças, que poderiam recomendar a aplicação de modelo geométrico sem tantas restrições.

MILLER (1991) já explorou um caminho muito tortuoso de pesquisa: o ***contexto***. RODRÍGUEZ (2000) admitiu ter sido a parte mais complexa de sua tese, quando estendeu a capacidade do seu MSS para lidar com informação contextual. RADA (1989), por seu turno, considerou a SS como variável dependente do contexto.

As representações de *contexto* estão muito ligadas ao *uso da língua* e é o tema de muitos estudos de lingüistas e psicólogos cognitivos, conforme frisou MILLER (1991). Para ele, contexto = conhecimento sobre uma palavra e sobre o uso dessa palavra. O autor distinguiu o contexto em sentido estrito (lingüístico) e amplo (aspectos lingüísticos e não-lingüísticos).

Esta tese não explorará o contexto na implementação dos dois protótipos. Por isso, na conclusão, trabalhos futuros serão incentivados nesse assunto. Cabe registrar que mesmo não se explorando o contexto nos instrumentos de avaliação de SS (os protótipos), de certa forma, este quesito foi contemplado de forma subliminar, como se explica em várias passagens desta tese.

JIANG (1997), RESNICK (1999), WONG (2000), SANTOS (2002) e GANESAN (2002) são todos autores de artigos científicos que contribuíram com trabalhos de natureza prática no campo de SRIs, cujo enfoque comum está na **SS nos domínios da Álgebra Linear**.

De uma forma ou de outra, esses autores se enquadram no que MEDEIROS (1999) compulsou como estado-da-arte em aplicações de SRIs (V. subitem 1.2.3), com apoio da IA. Em todos, o objetivo geral e comum era desenvolver estratégias de recuperar a informação contida em coleções de dados, ao medir-se a similaridade entre uma consulta e um (ou vários) documento(s) contido(s) nessas coleções.

Para o cálculo da SS[193], cada autor usou uma formulação especial derivada da noção do *produto interno* entre dois vetores, em que os vetores seriam os conjuntos de termos a serem comparados (termos da consulta x termos dos documentos).

De todos, extraiu-se uma conclusão fundamental, que de certa forma se coaduna com o que BÄHR (2000) havia registrado sobre o contexto e as relações entre entidades espaciais, que pode ser parafraseado na seguinte conclusão: "Se os *corpora* sobre os quais os programas de recuperação de informação forem aplicados estiverem previamente organizados, então haverá um ganho substancial nos resultados apresentados pela função de cálcu-

---

[193] A maioria não denominou de similaridade semântica (SS), mas apenas de similaridade.

lo da SS entre os termos apresentados na consulta com os termos que povoam a base de dados consultada".

Segundo esses autores que adotaram como modelo matemático de cálculo da SS o co-seno entre dois vetores num espaço euclidiano *n-dimensional*, cada vetor corresponde a um conjunto de termos da consulta ou a um conjunto de termos (documento) da base de dados consultada. A Equação 3.1 e a Figura 3.12 ilustram este enfoque.

$$sim(c_1, c_2) = \frac{\vec{c}_1 \cdot \vec{c}_2}{\|\vec{c}_1\| \times \|\vec{c}_2\|} = \frac{\sum_{i=1}^{n} \left(w_{1_i} \times w_{2_i}\right)}{\sqrt{\sum_{i=1}^{n} w_{1_i}^{\ 2}} \times \sqrt{\sum_{i=1}^{n} w_{2_i}^{\ 2}}} \qquad \text{Eq. 3.1}$$

A Equação 3.1 foi adaptada dos trabalhos de WONG (2000), SANTOS (2002) e GANESAN (2002). Este último autor a denominou de *co-seno de similaridade generalizada*.

A equação anterior pode ser interpretada geometricamente pelo ângulo que dois vetores fazem entre si, num espaço *n-dimensional*. Quanto maior o ângulo entre os vetores representativos das coleções de termos c_1 e c_2, menos similares são essas coleções[194]. No caso, as coleções foram substituídas não pelos termos, mas pelos pesos (w_i) que estes representam na operação de recuperação da informação. Na metodologia (Cap. 6), isto ficará mais claro.

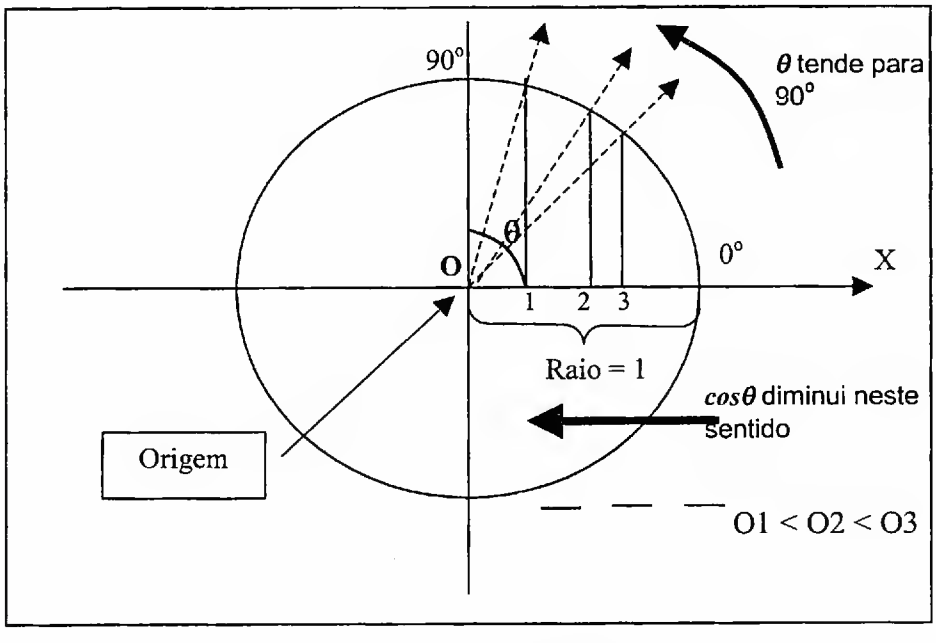

Figura 3.12: Interpretação geométrica da similaridade semântica.

[194] GANESAN (2002) denomina estas coleções de *bag of elements*.

A Figura 3.12 dá uma idéia da interpretação geométrica da SS. Nota-se que se o ângulo cresce, a similaridade decresce e o co-seno entre os dois vetores também decresce, já que no formalismo trigonométrico (círculo de raio unitário), o co-seno é medido no eixo ***X*** (abscissas) e, à proporção que o ângulo tende para 90°, no sentido anti-horário, é fácil verificar que o co-seno do ângulo θ tende para zero (o tamanho dos segmentos $\overline{O3}$, $\overline{O2}$ e $\overline{O1}$ diminuindo, à proporção que θ tende para 90°).

3.2.2.2.3. Tópicos sobre IA e OO – novos modos de representar a informação

IA e OO têm muito em comum. Mas como isso começou? E como isso está prosseguindo?

Para tentar visualizar o cenário pintado por essas duas perguntas, é preciso entender que o *software* é mais do que um diferenciador de produtos, sistemas e serviços, determinando vantagem competitiva no mercado. Os programas, documentos e dados que são o *software* ajudam a gerar comodidade e o mais importante fundamento de poder que qualquer pessoa, negócio ou governo podem adquirir: **INFORMAÇÃO**! (PRESSMAN,1995, p. 1.010).

Há cerca de 25 anos, o termo "processamento de dados" era usado para descrever o uso do computador num contexto comercial (máquina de calcular e memória). Hoje, o termo evoluiu para "tecnologia da informação (TI)", que abrange a acepção anterior, mas muda o foco para extrair informação significativa de dados (V. "dado, informação, conhecimento e sabedoria" no glossário).

A mudança nas TIs que exercem um impacto sobre a computação parece assumir uma progressão estabelecida empiricamente por Michael Horner [*apud* PRESSMAN (1995, p.1.011)] pela chamada ***regra 5-5-5***. Essa regra quer dizer o seguinte: um conceito essencialmente novo parece deslocar-se da idéia inicial para um produto de mercado em cerca de quinze anos. São cinco anos de idéia, mais cinco anos de prototipação[195] e mais cinco de testes do lote-piloto para entrar em produção.

Os novos modos de representar informação podem ser compreendidos, ao se fazer uma simples avaliação da recente história da computação. Nas últimas quatro décadas, as mudanças sofridas no desenvolvimento de *hardware*[196] e *software* foram impulsionadas pelos avanços nas ciências "hard" ou exatas.

---

[195] V. protótipo no subitem 6.3.1.

[196] Mais nessa parte, pois as mudanças de *software* eram muito dependentes do *hardware*.

A regra 5-5-5 parece funcionar razoavelmente bem quando novas tecnologias são derivadas das ciências exatas. Contudo, já há indícios que para as próximas décadas os avanços da computação receberão impulso das chamadas ciências "soft" aplicadas, como a Psicologia, a Sociologia, a Ciência da Informação e outras. PRESSMAN (1995) admitiu que é muito difícil prever como será a gênese dessas novas mudanças, dando como exemplo, justamente, o caso da inteligência humana, que vem sendo estudada há séculos sem que haja ainda uma compreensão a respeito de sua natureza.

O que é inquestionável é que a influência das ciências sociais e humanas já começa a afetar os projetos de computadores do futuro, até hoje mais orientados pela microeletrônica tradicional do que pela compreensão da fisiologia cerebral e dos processos mentais. Esse enfoque é sustentado por quatro elementos: 1) As pessoas que fazem o trabalho, 2) O processo que elas aplicam, 3) A natureza da informação e 4) As TIs subjacentes.

É sobre o quarto elemento acima que se aplica o assunto em foco: IA e OO. Para se avaliar as mudanças por que passará o insumo que alimentará esses tipos de *software*, é preciso avaliar as tendências de *hardware*, que sempre determinaram e continuarão a determinar o *software*.

São dois os caminhos que se prognosticam para o *hardware*. Um deles será a continuidade da evolução das tecnologias de *hardware* já amadurecidas, baseadas na microeletrônica tradicional, responsável por um sem-número de aplicações de baixa a média complexidade, mas que envolvem colossais quantidades de dados e muito processamento. Nesse domínio estão as arquiteturas CISC (*complex instruction set computer*) e RISC (*reduced instruction set computer*), com seus nichos[197] de aplicações correspondentes.

O outro caminho, que começa a se configurar, vem do desenvolvimento de arquiteturas de *hardware* não tradicionais, p.ex: máquinas maciçamente paralelas (algumas até de fundamentação RISC), processadores ópticos e máquinas de redes neurais, já em fase de prototipação (segundo "5" da regra 5-5-5).

De fato, as tecnologias OO podem formar uma ponte entre os enfoques da IA (por enquanto, inerentemente OO), as aplicações de *software* convencionais e a tecnologia de bancos de dados. Ao fazer isso, elas podem representar um passo importante na direção do processamento do conhecimento [(PRESSMAN, 1995) e (RODRÍGUEZ, 2000, p. 48)].

[197] Em tese, máquinas CISC tendem para plataformas domésticas e medianamente corporativas. As RISC, para plataformas pesadamente corporativas (pelo seu alto desempenho).

É possível que outros enfoques mais adequados para a implementação de problemas da IA venham a surgir, mas a OO, por enquanto, é o mais adequado, porque "enxerga" o mundo real de uma forma mais simples do que a da clássica análise estruturada.

Trabalhos como o de SILVA (2001, p. 80 e 128) identificaram várias correspondências entre o desenvolvimento das metodologias OO e as metodologias adequadas à IA e pertencentes às ciências da classificação (Terminologia, p.ex.), que sempre operaram na Teoria do Conceito. Este autor tabulou várias dessas correspondências conceituais, como p.ex: ***conceito*** da Terminologia e ***classe*** da OO; ***características*** da Terminologia e ***atributos*** e operações de classe na OO; ***termo*** da Terminologia e o ***nome da classe*** na OO; o ***referente*** da Terminologia e o ***objeto*** na OO; a ***extensão*** da Terminologia e ***objetos*** na OO.

Outros autores, como CRANEFIELD (2002), identificaram os requisitos de modelagem conceitual para construir ontologias[198] de sistemas de catalogação, em que a UML™ foi eleita como padrão de modelagem para a ulterior implementação (por meio de linguagens hospedeiras que a UML™ possui para gerar código: *Java* ™, p.ex.).

Uma das características da OO que mais pode contribuir para acelerar a construção de AIs é a *reutilização* de *software* (DEITEL, 2000). Uma verdadeira indústria já surgiu, salientando-se os seguintes atores: construtores de *dispositivos reusáveis de software*; construtores de dispositivos para sistemas (HCIs) e integradores das duas soluções anteriores, para atender às necessidades específicas de usuários (PRESSMAN, 1995).

No que tange às ferramentas de análise que poderão ser requeridas durante a metodologia de construção dos protótipos desta pesquisa, já é possível adiantar que, como se tratam de protótipos, é possível deixar de lado a apresentação circunstanciada das notações de projeto. Nesse caso, a documentação (completa e precisa) dos programas já é suficiente.

O material relacionado acima já supriria a necessidade de entendimento do PRONTO©. No entanto, como se trata de um estudo exploratório de manipulação experimental, é de bom alvitre suprir os possíveis futuros sucessores dessa linha de pesquisa com mais material documentário, a fim de que possam retomar a intenção de construir uma base de conhecimentos cartográficos de um ponto mais distante que a origem.

Por isso, além dos esforços para documentar o código-fonte, o PRONTO© foi contemplado com uma "roupagem" parcialmente sistêmica, juntando-se duas notações gráficas: **o diagrama estático de classes**, que denota o *invariante* estrutural do sistema[199] e o **dia-**

[198] Corpo de conhecimentos de um domínio específico, formalmente modelado, para definir conceitos e estabelecer relações entre eles (CRANEFIELD, 2002).
[199] Pelo menos em estado embrionário no PRONTO©.

grama de caso de uso, que descreve as funções do sistema e suas fronteiras, percebidas por atores externos.

Os dois diagramas citados são simples, básicos e universais (unificados), visto como reproduzem a visão geral das classes de entidades e seus principais relacionamentos. Essas três características dessa notação foram adaptadas ao corpo de uma linguagem unificada de modelagem, denominada UML™ (*Unified Modeling Language*).

A UML™ é uma linguagem padronizada, composta de um conjunto de procedimentos e notações para especificar, visualizar, documentar e construir produtos de programação de um sistema, para ser utilizada em todos os processos ao longo do ciclo de desenvolvimento de sistemas orientados a objeto ou por diferentes tecnologias de implementação (FURLAN, 1998, p. 33).

Da definição, conclui-se que a UML™ é uma linguagem e, portanto, não deve ser confundida com os métodos orientados a objeto (V. glossário). Pode ser aplicada a qualquer método orientado a objeto e é incompleta como método, porque não possui a descrição de um processo-guia do desenvolvimento de sistemas.

A UML™ apenas fornece as notações formais para a reprodução dos diagramas de modelagem de sistemas orientados a objeto. Enquanto que um método, além da notação, possui uma seqüência de tarefas a serem seguidas para construir um sistema.

A gênese de um sistema de informação está na fase da *análise*. Segundo SILVA (2001), é nessa fase que a representação e a estruturação das informações indicam a possível aplicação da Terminologia e, por extensão, o emprego dos conhecimentos advindos da área das ciências da classificação (Biblioteconomia e Ciência da Informação). A semelhança dos conceitos usados na UML™ e aqueles empregados na construção de *tesauros* é uma prova disso, segundo o mesmo autor, que desenvolveu um estudo-de-caso comparativo entre esses dois tipos de linguagens documentárias.

Um trabalho muito importante de cobertura das lacunas entre as pesquisas de linguagens de desenvolvimento de sistemas, na Ciência da Computação, e o desenvolvimento de *tesauros*, na Ciência da Informação, foi a dissertação de SILVA (2001), comparando de diversas formas esses dois tipos de ferramentas.

Mesmo sem o saber, SILVA (2001) fornecia parte da resposta para as inquietantes questões de SÖRGEL (1999) em relação à falta de sinergia entre os dois campos do conhecimento científico envolvidos com tais ferramentas (*tesauro* e UML™).

O enfoque interessante do autor é o de trazer para a análise as diferentes formas de apresentação dos mesmos conceitos com que lidam a Terminologia e a Ciência da Computação, especificamente, a OO.

Os conceitos básicos da OO, que se projetam na Terminologia, são os de *objeto* e *classe*. Segundo DEITEL (2000, p.77 e 142), um **objeto** é um componente de *software*, essencialmente reutilizável, que modela uma parcela do mundo-real. Percebe-se a estreita ligação de objeto com o referente do triângulo do conceito de I. Dahlberg (V. Figura 3.7). Assim como sem objeto não existe programa, na POO, da mesma forma, sem o referente não existe conceito na Terminologia. Essa característica essencial da Teoria do Conceito e da TMO foi incorporada como um dos mais importantes axiomas do PRONTO© (Axioma 15, subitem 6.3.1.2).

Já as **classes** são estruturas capazes de gerar e conter objetos (instâncias da classe) com características semelhantes.

Quadro 3.4: Comparação entre conceitos da OO e da Terminologia[200].

TEORIA DO CONCEITO (Terminologia)	**ORIENTAÇÃO A OBJETO** (Ciência da Computação)
Conceito	Classe
Características	Atributos e operações da classe
Termo	Nome da classe
Extensão	Objetos

Não constitui um objetivo desta tese estudar a fundo a UML™ e nem tampouco um *tesauro*, mas como foi o enfoque da OO que fundamentou a construção da folha Faxinal© (*corpus* para o PRONTO©), como já se demonstrou a importância dos métodos de construção de linguagens documentárias para subsidiar a montagem de ontologias, e como o trabalho de SILVA (2001, 118-154) operou nessa zona de limbo, vale a pena visualizar as similaridades de conceitos que este autor ilustrou e que são reproduzidas no Quadro 3.4.

[200] Adaptado de SILVA (2001).

O que SILVA (2001) concluiu de relevância para esta pesquisa, das comparações que efetuou entre um *tesauro* e a UML™, no âmbito do desenvolvimento de sistemas de informação, pode ser sintetizado nos seguintes tópicos:

- O enfoque no referente (real ou abstrato) do conceito (V. Figura 3.7) permite a identificação de características que aumentam a compreensão desse referente (entidade), o que é básico na análise das informações na UML™;
- A identificação de mais opções de relacionamentos entre termos, como todo-parte no espaço e no tempo (relações de agregação na OO), permite enriquecer o conteúdo dos conceitos na UML™;
- Apesar da reduzida possibilidade do emprego direto do *tesauro* na construção de bases de dados orientadas a objeto, é possível utilizar as especificações de construção de um *tesauro* e os princípios da Terminologia na estruturação da UML™, particularmente na representação da informação;
- Verifica-se, então, a ampliação do uso dos elementos do *tesauro* na UML™. Assim, de uma forma geral, a UML™ é constituída de procedimentos que podem empregar os princípios da Terminologia na orientação a objeto;
- O ponto forte dos *tesauros* é a sua capacidade de representação e estruturação de informação fracamente formalizada;
- Já o ponto forte da UML™ é a sua capacidade para ser utilizada na modelagem orientada a objeto;
- A UML™ tem potencial para a construção de *tesauros* mais específicos;
- Diante de todos esses aspectos, o autor concluiu, finalmente, que há uma grande afinidade entre essas duas linguagens documentárias, uma complementando a outra em suas lacunas conceituais ou metodológicas.

Além de todas as capacidades fundamentais de modelagem de sistemas de que dispõe a UML™, ela também possui comodidades para gerar código-fonte em linguagens de programação (como o *Java*™, p.ex.), como saída para as suas notações gráficas.

No que tange à implementação dos protótipos, pretende-se utilizar a linguagem *Java* ™, que por ser OO pode reproduzir com mais simplicidade mecanismos de abstração da mente humana ao modelar a realidade por uma hierarquia de classes e também por ser capaz de incorporar mais semântica em suas construções do que as linguagens convencionais de programação.

Possuir um mecanismo de reprodução das relações de generalização e especialização de classes de entidades é uma característica de fundamental importância para a utilização

de uma linguagem OO como o *Java*™. As linguagens apenas estruturadas não incorporam esse mecanismo. Em geral, segundo FURLAN (1998, p.60), as LTPs apenas suportam mecanismos de reprodução das relações meronímicas (todo-parte). Aí está uma forte razão pela escolha do *Java*™ para a implementação dos protótipos.

As origens do *Java*™ remontam ao Projeto GREEN, nos laboratórios da Sun Microsystems, onde começou com o nome de Oak (por causa do *carvalho* que entrava pela janela do chefe do projeto, James Gosling), passou incidentalmente para "Java" por uma decisão do grupo de pesquisa que tomava café num botequim local. Este café tinha o nome de "Java" (DEITEL, 2000).

Hoje em dia, o *Java*™ está em vertiginoso crescimento em todo o mundo da Informática. Na verdade, está ocorrendo um deslocamento significativo de opção entre os programadores que vieram do tradicional Smalltalk™[201] e até mesmo do C++™ (FURLAN, 1998).

O *Java*™ revolucionou o desenvolvimento de *software* para quase todos os ambientes distribuídos (redes). Ela se caracteriza por ser uma poderosa linguagem OO, dinâmica, independente de plataforma tecnológica, segura e relativamente fácil de usar. Sua sintaxe[202] é muito parecida com a do C++™, favorecendo aqueles que já tiveram familiaridade com esta LTP; a diferença é que o código em *Java*™ não precisa ser compilado, o que acelera o processo de desenvolvimento em relação ao C++™ (FURLAN, 1998).

Outra característica de relevo está centrada no rigor das especificações de seus tipos de dados, recebendo-a de suas tributárias, o C e o C++ (DEITEL, 2001, p. 177). Por ser fortemente *tipada*, sua estrutura sintática se ajusta muito bem ao formalismo das notações em BNF para as ontologias sobre as quais o PRONTO© calculará a SS, como se verá no Cap. 6.

Por todas as características já descritas das linguagens OO, particularmente por serem as mais adequadas para implementar os complexos problemas de IA, a LTP *Java*™ foi a escolhida por uma virtual vantagem em relação às suas próprias congêneres: o benefício da reutilização de código é multiplicado pela existência de uma extensa indústria de *software* que lhe dá suporte. É muito fácil encontrar na Internet bibliotecas de códigos e classes de objetos feitos sob medida para uma aplicação particular, sem qualquer ônus para o usuário-programador.

Em que pese a adequação das ferramentas de modelagens OO como a UML™ e as linguagens de programação OO participarem de um significante papel na formalização de problemas da IA, um novo campo das engenharias que lidam com dados altamente estrutu-

---

[201] Desenvolvida pelo *Learning Research Group* do Palo Alto Research Center (PARC) da Xerox (DEITEL, 2000).
[202] Eis um outro motivo da escolha por esta linguagem, já que o pesquisador tem relativo conhecimento do C++™.

rados (informação e conhecimento) está emergindo. Trata-se da Engenharia de Ontologias [(DIAS, 2003), (BRANDÃO, 2003) e (GUIZZARDI, 2003), cujo objeto, ainda que nebuloso em seus primórdios, parece se delinear na forma de um conjunto de técnicas e de conceitos colocados à disposição de profissionais e pesquisadores interessados no desenvolvimento de linguagens destinadas à modelagem rápida e sistematizada de ontologias e no uso dessas linguagens ou aplicativos delas derivados, como novos paradigmas no ciclo de vida de sistemas do século XXI.

Esses tipos de linguagens e de aplicativos são espécies de linguagens de representação do conhecimento (LRC), que podem ser distinguidas como LRO (linguagens de representação de ontologias), muito mais ajustadas ao conceito mais difundido de ontologia, quando se trata de desenvolver sistemas de informação, qual seja:

"Ontologias são modelos de um domínio restrito do conhecimento, formalmente especificadas e destinadas a definir os conceitos e suas relações nesse domínio."

Nesse ponto, é preciso distinguir dois ramos de LROs para formalizar ontologias: o ramo dos aplicativos ou de linguagens específicas de domínio público e as linguagens de fundamentação OO e de propósito geral.

O primeiro ramo acima citado engloba linguagens de estrutura sintática linear (linha de comando), bem assemelhadas às LTPs construídas para declarar sentenças lógicas da LPO (PROLOG) ou para manipular símbolos (LISP) [203].

O segundo ramo é representado pela UML™, que é uma linguagem padronizada, composta de um conjunto de procedimentos e notações para especificar, visualizar, documentar e construir produtos de programação de um sistema e para ser utilizada em todos os processos ao longo do ciclo de desenvolvimento de sistemas orientados a objeto ou por diferentes tecnologias de implementação (FURLAN, 1998, p. 33).

A vantagem das linguagens do primeiro ramo está no seu alto poder de expressividade de enunciados da LPO. Em contrapartida, são de conhecimento restrito aos programadores e pesquisadores de IA, de difícil aprendizado e, por conseguinte, não se expandem muito além dos ambientes acadêmicos, em que pese os esforços de desenvolvimento de interfaces gráficas de edição de ontologias para aumentar o nível de abstração e, conseqüentemente, para incrementar o poder de representação dessas linguagens.

No caso de uma linguagem de modelagem como a UML™ (segundo ramo), as vantagens merecem ser listadas:

---

[203] LISP™ (criada por John McCarthy, no MIT, em 1958) e PROLOG™ (criada por Alain Colmerauer, na Universidade de Aix-Marseille – França -, em 1972).

- Expressiva penetração dessa tecnologia em diversos ambientes, além do puramente acadêmico;
- As LTPs OO que implementam os seus modelos possuem uma vasta biblioteca (repertório) de objetos que podem ser combinados para atender às mais diversificadas necessidades, (SILVA, 2001);
- Alto poder de representatividade para problemas complexos.

No primeiro ramo, a literatura vem expondo várias iniciativas de natureza prototipada. Entre elas, salientam-se o KIF™ [*knowledge Interchange Format - National Committee for Information Technology Standards*, 1998, *apud* CRANEFIELD (2002)] e o KL-ONE™ [R. J. Branchman e J. G. Schnolze, 1985, *ib.*]. Ambas são linguagens de alta expressividade, mas de baixo nível, no que tange à representação gráfica de ontologias (nesse caso, a UML™ é mais adequada).

Esforços têm sido feitos para desenvolver ferramentas gráficas de edição de ontologias, como os do Laboratório de Conhecimento Compartilhado da Universidade de Stanford, que derivou um subconjunto do KIF™, denominado de *Ontolingua*™, que oferece comodidade na representação gráfica de ontologias em alto nível de abstração.

O OMG também se preocupa com o estabelecimento de especificações para o desenvolvimento de ontologias. Trata-se do padrão XMI™ (*XML Model Interchange*), que já foi publicado e vem recebendo contínuos aprimoramentos para incrementar o poder de representatividade da UML™ nesse campo.

O que se deve ter mente na concepção de uma estrutura de representação do conhecimento, como uma LRO, é a noção da função que a semântica exerce no campo da IA. Ela serve como uma camada de mediação (uma "ponte") entre os fatos do MR e as sentenças do universo lingüístico que se referem àqueles fatos. Aos fatos, sucedem-se outros fatos e a sua tradução por sentenças no universo lingüístico deve assegurar, o mais fidedignamente possível, a representação desta sucessão de fatos do MR.

Os modelos descritivos em LN têm sido, por muito tempo, o meio pelo qual se manifestou a semântica na tarefa de manter um paralelismo entre os fatos que ocorrem no MR e as sentenças expressas no mundo lingüístico, mas com o advento dos sistemas computacionais, foi necessário criar um novo meio de transladar as sentenças desse último mundo para um ambiente de *bits* e *bytes*.

Surgem, então, na década de 50 do séc. XX, as primeiras LTPs. O problema com essas linguagens é que não houve preocupação em projetá-las para representar os fatos do mundo como eles são ou como, pelo menos, poderiam ser. As LTPs, em geral (FORTRAN,

COBOL, PASCAL, C, BASIC, etc), foram criadas mais para descrever o estado interno das máquinas e como ele evolui no tempo, à proporção que um programa é executado.

À propriedade de uma LTP de representar o mais fielmente possível o MR denomina-se ***expressividade***; e boa parte delas é pouco expressiva, ao contrário das LNs (português, sueco, etc.), que têm compromissos maiores com a comunicação entre os seres humanos e, portanto, são pouco representativas, dependendo muito do contexto em que se manifestam, além de contar com estruturas (frases) que não trazem em si (portam) conhecimento explícito, o que as torna muito susceptíveis ao fenômeno da ambigüidade.

Idealmente, uma boa LRC deveria incorporar as vantagens de *representação* das LTPs e de *expressividade* das LNs, propiciando aos seres humanos comunicarem-se de forma não-ambígua[204], i.e., de forma clara, precisa e concisa [(SALMON, 1973, p.137) e RUSSELL, 1995, p. 680-682 e p. 712-715]].

A base da IA, e de qualquer sistema que sobre os seus fundamentos se apóie, repousa numa LRC de base lógica, não importando qual o tipo de notação (codificação) que será adotada. O que importa é que, com ela, seja possível construir uma arquitetura de dados altamente estruturados (conhecimento) e, também, que seja possível extrair informações dessa estrutura, em ambos os casos, de maneira simples e elegante, por um mecanismo que simule os processos de raciocínio humano.

Um argumento só pode "capturar" significado, se ligar o enunciado escrito a um fato do mundo (isto é semântica). Para isso, é preciso que o criador do argumento inclua nele *interpretação.* O problema é que os seres humanos não estão acostumados (e nem o poderiam por questões evolucionárias) a explicitar interpretação nas sentenças que expressam em LN. O que fazem, naturalmente, no momento de se comunicarem entre si, é encobrir ou dissimular (metáforas) a interpretação nas suas construções morfossintáticas.

Portanto, desenvolver uma LRC é uma tarefa nada trivial. A interpretação que se vai dar às sentenças (argumentos) deve seguir uma composição sistemática, i.e., o significado do todo deve estar explicitamente ligado aos das partes, como num silogismo. A métrica máxima de um AI deve ser adequada à construção de uma LRC, ou seja: "A validade do argumento depende, então, tanto da interpretação da sentença que o forma como do atual estado do mundo que ele representa".

Todas as considerações anteriores sobre os problemas ligados à criação de uma LRO já desencadearam iniciativas para definir uma nova área de estudos dentro da emergente

---

[204] Que isto fique apenas no âmbito da IA, pois o que seria da prosa, da poesia, enfim, da estilística em geral?

Engenharia de Ontologias. Segundo GUIZZARDI (2003), trata-se da ***Análise de Domínio***, considerada importante pelos engenheiros de *software*, porque, entre outras razões, atenderia à necessidade de diminuir o custo desproporcional da manutenção de *software* resultante da introdução de mudanças arbitrárias, assim como do consenso relacionado à importância do desenvolvimento para a finalidade de reutilização de *software*.

Assim, a Análise de Domínio pode ser considerada como uma área de estudos que procura identificar os objetos, operações e relações entre o que peritos (especialistas) num determinado domínio do conhecimento percebem como importante (GUIZZARDI, 2003).

W. J. Clancey [*apud* GUIZZARDI (2003)] propôs um foco alternativo para a Engenharia de Conhecimento, que deveria calcar-se na modelagem de sistemas e não na tentativa de reproduzir a maneira como os especialistas raciocinam. Segundo o autor, uma base de conhecimentos deve ser vista como um produto de uma atividade de modelagem e não um repositório de conhecimento especializado. Dessa forma, a modelagem passa a ser o aspecto central da Engenharia de Conhecimento e a aquisição de conhecimento passa a ser essencialmente um processo construtivo, no qual o engenheiro de conhecimento usa todos os tipos de informação disponíveis e estabelece as decisões finais de modelagem.

Dentro da comunidade de representação do conhecimento surgiu, então, a idéia de que o conhecimento embutido em uma determinada porção da realidade poderia (e deveria) ser representado em um nível de abstração, tal que fosse independente e reutilizável ao longo de várias tarefas. Ao adotar esse paradigma, essa comunidade entrou em um território anteriormente explorado unicamente por filósofos da ciência e da linguagem, acelerando e aprofundando as investigações que os filósofos e lingüistas até então realizavam. Ao resultado dessa interdisciplinaridade sobre um problema inicialmente criado por Aristóteles, abrangendo estudos de classificação, sistemas de taxinomia e de representação do conhecimento, de forma geral, denomina-se, hoje, de Engenharia de Ontologias (GUIZZARDI, 2003).

Ampliando as definições anteriores sobre o objeto dessa nova engenharia – a ontologia -, é instrutivo finalizar a revisão de literatura, examinando alguns aspectos suplementares (**definições específicas**, **classificação**, **métricas de construção** e **uso**) dessa nova estrutura de representação do conhecimento.

Segundo GUIZZARDI (2003), as dificuldades atuais enfrentadas nessa área de estudos das ontologias são de caráter metodológico. Apesar de uma grande quantidade de ontologias já ter sido desenvolvida por diferentes grupos, por diferentes enfoques e utilizando diferentes métodos e técnicas, poucos trabalhos foram publicados sobre como proceder, mostrando as práticas, critérios de projeto, atividades, métodos e ferramentas empregadas para

a construção dessas estruturas. A conseqüência é clara: inexistência de ciclos de vida, métodos sistemáticos, critérios de qualidade, técnicas e ferramentas, expondo o desenvolvimento de ontologias aos mesmos problemas vivenciados nos primórdios da Engenharia de *Sofware*, há mais de trinta anos atrás, quando os seus problemas eram mais resolvidos pela via improvisada do ensaio-e-erro do que pelo rigor do método das engenharias.

Algumas propostas de metodologia para a construção de ontologias têm sido apresentadas na literatura nos últimos anos. GUIZZARDI (2003) e LEÃO (2003) relataram algumas iniciativas nesse sentido, tanto em LROs como em projetos de sistemas com fundamentação em ontologias. No primeiro caso, além do KIF™, KL-ONE™ e *Ontolingua*™, já citadas pelo primeiro autor, seguem-se as citadas pelo segundo autor:

- CML™ (*Conceptual Modelling Language*), de G. Schreiber *et al.*;
- SHOE™ (*Simple HTML Ontology Extension*), de J. Heflin;
- OIL™ (*Ontology Interchange Language*), de I. Horrocks *et al.*;
- IDL™ (*Interface Definition Language*), de T. J. Mowbray e R. Zahavi;
- TOL™ (*Task Ontology representation Language*), de M. Ikeda *et al.*;
- FLogic™, de M. Kifer *et al.*;
- LINGO™ (LINguagem Gráfica para descrever Ontologias), de R. A. Falbo e G. Guizzardi.

No segundo caso, LEÃO (2003) e GUIZZARDI (2003) citaram os seguintes projetos:

- ESPRIT Kactus™, de G. Schreiber *et al.*;
- DARPA™ (*Joint Forces Air Component Commander*), de A. Valente *et al.*;
- CYC™, de D. Lenat e R. Guha;
- TOVE™ (*Toronto Virtual Enterprise*), de M. Fernandez *et al.*;
- *Methontology*™ de M. Uschold e M. Grüninger.

Apesar dos esforços, os modelos apresentados ainda não produzem resultados de ordem funcional suficientemente satisfatórios a ponto de suportar a construção de ontologias como uma verdadeira disciplina de engenharia. O que existe é alguma orientação e sistematização do conhecimento já produzido nesse campo.

As **definições específicas** a seguir são fruto tanto das reflexões de ordem filosófica e lingüística que se fazem sobre o assunto (*ontologias*), bem como sobre as avaliações que se tem levantado sobre os resultados apresentados por diversos esforços de pesquisa experimentais.

Para RODRÍGUEZ (2000), uma ontologia é um tipo de BC que descreve os conceitos de um determinado domínio do saber, por meio de definições suficientemente pormenorizadas, capazes de capturar a semântica de uma certa visão do mundo e de permitir consultas

intensionais ao conteúdo informativo de bases de dados, i.e., consultas que recuperem informação relevante dessas bases, sem levar em conta a representação das camadas de dados (estruturas de dados).

Para MEDEIROS (1999), uma ontologia é uma rede ou malha (*treillis*) de conceitos, em que as restrições (regras) semânticas estabelecem o mecanismo de herança na malha, ou seja, a carga de subtipos na estrutura só se efetiva se estes herdam as características de seus supertipos.

Para USCHOLD (1996), uma ontologia pode assumir uma variedade de formas, mas necessariamente deve incluir um conjunto de termos específicos. Uma ontologia é virtualmente uma manifestação de um entendimento compartilhado que se tem sobre um domínio do conhecimento. Essa concordância no entendimento propicia uma comunicação mais precisa entre as partes interessadas nesse domínio, o que acaba por produzir benefícios nos campos da interoperabilidade e da reutilização de idéias.

E as definições de *conceitualização*, relevância e representação?

Ainda segundo USCHOLD (1996), *conceitualização* é uma visão do mundo; um conjunto de regras informais que descrevem uma porção do MR; um sistema de conceitos (entidades, processos e atributos), suas definições e suas relações entre si. Esse processo funcional da mente é bem similar à noção de *esquema* [V. MORA (1994), p. 146].

Segundo JOHN (1994), a relevância é uma função que determina se um conjunto de feições conceituais, armazenadas numa base de dados altamente estruturados (BC), é ou não (irrelevante) capaz de classificar novos conceitos que sejam apresentados para o processo de incorporação a esta BC.

O conceito de *representação*, segundo MORA (1994), está ligado aos de *esquema* e de *imagem* ou *representação mental* de Kant. A idéia kantiana para *esquema* era a de uma representação homogênea e mediadora[205] entre a categoria (idéia) e a aparência (forma) do fenômeno, tal que fosse possível a aplicação da primeira (categoria) à segunda (aparência). Como se viu antes, o esquema possui uma natureza introspectiva, enquanto a representação mental necessita de estimulação externa (mundo empírico) para se realizar.

Ainda segundo JOHN (1994), a *representação* é um fenômeno associado ao critério de julgamento de similaridade entre entidades do mundo, que partilhem de características comuns com uma categoria de classes mais genérica [V. MORA (1994), p. 146].

---

[205] Mais uma vez uma semelhança com os três mundos *popperianos*.

TVERSKY (1977) considerava a similaridade como uma variável independente de uma relação de representação.

Quase metaforicamente, HENRIQUES (2002) reflete sobre esses conceitos acessórios para a compreensão do que seja ontologia, comparando a mente a uma máquina abstrata que manipula símbolos, os quais "tiram" seu sentido de uma correspondência (representação) que a mente (mundo interior) faz com as coisas do mundo exterior. A mente, para o autor, é um espelho da natureza, em que a razão da primeira espelha (representa) corretamente a lógica da segunda.

Há dois enfoques de **classificação** de ontologias: o de G. van Heijst e o de N. Guarino [*apud* LEÃO (2003)].

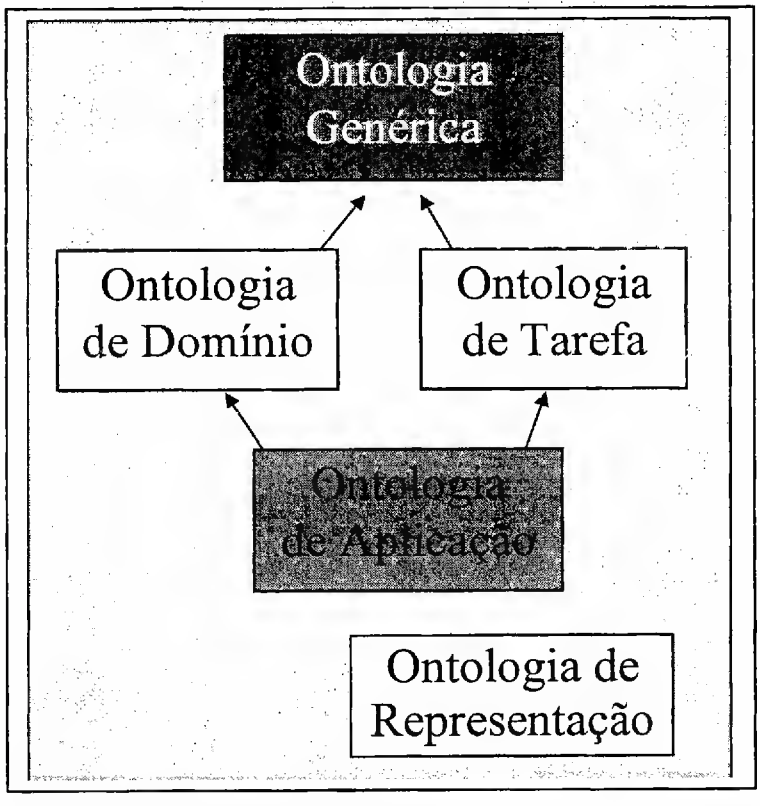

Figura 3.13: Níveis de generalização das ontologias[206].

Van Heijst adotou o critério de classificar as ontologias quanto ao tipo de estrutura e quanto ao assunto da conceituação. Pelo primeiro critério, as ontologias podem ser:

- Terminológicas;
- De informação;

---

[206] Retirada de LEÃO (2003).

- De modelagem do conhecimento.

Pelo segundo critério de van Heijst, partilhado por N. Guarino, as ontologias podem ser:

- Genéricas;
- De domínio;
- De tarefas;
- De aplicação;
- De representação.

A Figura 3.13 ilustra a visão de N. Guarino sobre os níveis de especialização/generalização das ontologias. As que estão sendo apontadas pelas setas são as mais genéricas, ou seja, as ontologias de domínio e de tarefa contêm termos que descrevem os conceitos e relações existentes no nível das genéricas.

As *genéricas* descrevem conceitos muito gerais, tais como: espaço, tempo, matéria, objeto, evento, ação, etc., que são independentes de um problema ou domínio particular. Essas ontologias se destinam a criar teorias básicas do mundo, típicas das áreas de categorização filosófica e de Lingüística (p.ex: CYC™ e Wordnet™).

As *de domínio*, mais comuns, expressam conceitos de domínios particulares, ao pormenorizar o vocabulário de um domínio genérico (p.ex: TOVE™, MSS e PRONTO©).

As *de tarefas* expressam conceitos ligados à resolução de problemas, independentemente do domínio em que ocorram, constituindo-se, portanto, em elementos de integração entre distintos domínios.

As *de aplicação* descrevem conceitos dependentes de domínio e tarefa particulares.

As de representação explicam os conceitos que fundamentam a formalização do conhecimento (p.ex: *Ontolingua*™).

Quanto às **métricas gerais para o desenvolvimento de ontologias**, USCHOLD (1996) indicou os seguintes passos:

- **Construção da Ontologia:** *1) Captura* - identificação dos conceitos e relações relevantes no domínio de interesse; *2) Codificação* - a conceituação capturada é representada em alguma linguagem formal; *3) Integração com ontologias existentes*.
- **Identificação do Propósito:** é a razão de construção de uma ontologia, seus usos projetados e os seus potenciais usuários, definindo qual será a sua utilização (compartilhamento, reutilização, organização e estruturação do conhecimento).
- **Avaliação:** em termos de questões de pertinência com relação ao domínio explicitado e especificações de requisitos.

- **Documentação:** todas as decisões importantes devem ser documentadas, bem como os conceitos e axiomas identificados.

Finalmente, as três grandes áreas de aplicação (**uso**) das ontologias, segundo M. Uschold, M. Grüninger e I. Kalfoglou [*apud* LEÃO (2003)], são as seguintes:

- **Comunicação:** entre pessoas e organizações, de forma menos ambígua;
- **Interoperabilidade:** entre SIs, como um formato de intercâmbio;
- **Engenharia de Sistemas:** na especificação de requisitos para um SI, na aquisição de conhecimento, na produção de *software* confiável (representação formal ⇒ automação confiável na fase de testes) e na reutilização de componentes.

3.2.3. Conclusão parcial da revisão de literatura da área de ciências cognitivas

Como Dagobert Sörgel (SÖRGEL, 1999) argumentou, muito esforço foi desperdiçado pelos pesquisadores das ciências exatas na área da Ciência da Computação, particularmente, em virtude da falta de esforços conjugados (interdisciplinaridade) com os pesquisadores das ciências sociais aplicadas e humanas no desenvolvimento de sistemas de informação. Pelo menos nesta revisão de literatura, não se levantou nenhum material que indicasse haver alguma relação entra a Teoria Geral da Lingüística de Saussure e a TMO de J. Rumbaugh ou a UML™.

Esta pesquisa é uma continuação de esforços para reverter o quadro pintado por SÖRGEL (1999).

O ponto comum da revisão de material das *geociências*, Ciência da Computação (IA) e outras ciências cognitivas é a tendência de convergência das pesquisas. Uma terminologia comum está gradualmente se formando. Não é raro verificar, p.ex., na trilogia de CÂMARA (2002), conceitos e princípios da área da Psicologia Cognitiva (a Lei de Tobler). Também não é raro o uso de metáforas computacionais pelos psicólogos cognitivos, o que vem ocorrendo desde a década de 60, com o trabalho de M. R. Quillian (Memória Semântica).

Dos vários modelos de avaliação de SS, percebe-se não mais haver um divisor d'água entre os modelos baseados na Teoria dos Conjuntos (A. Tversky), que contemplam mais a contagem ou freqüência de feições num domínio específico de estímulos e os modelos baseados na Geometria Analítica e Álgebra Linear (Roy Rada e posteriores), que traduzem os termos em vetores contidos num espaço geométrico euclidiano, para mensurar a SS pelo afastamento (*dissimilaridade*) ou pela aproximação (similaridade) desses vetores que representam os termos.

Neste ponto da pesquisa, parece que o caminho da avaliação da SS pelo enfoque de RODRÍGUEZ (2000), de caráter semântico, com todas as limitações que serão enunciadas na metodologia, é o mais viável para ser implementado e constituir um mínimo de evidência para a hipótese de pesquisa que se pretende formular. É desejável que esta hipótese (ou pressuposto) indique um tipo de formalismo, materializado num programa de computador, capaz de se aproximar da capacidade cognitiva de julgamento de semelhanças e de diferenças entre classes de entidades espaciais.

O MSS e o PRONTO© foram implementados por LTPs OO (o primeiro em C++™ e o segundo em *Java*™). Essa iniciativa de adoção de uma LTP OO para resolver o problema de transformar a explicitação de um sistema de conceitos (ontologia) por meio de linguagens de domínio público vem da ainda incipiente situação em que se encontra a Engenharia de *Software* nessa área.

Tais protótipos contam pontos para a meta de se instituir um campo da engenharia (Engenharia de Ontologias) que cuide das especificações de modelagem e projeto de ferramentas de desenvolvimento de sistemas de informação, de fundamentação ontológica.

Os dois protótipos citados se alinham com os esforços de instituições de ensino e grupos de pesquisa e de normalização internacionais - Universidade de Stanford e OMG, respectivamente -, pioneiras nesse tema.

Assim como o MSS, o PRONTO© deve ser capaz de automatizar os processos de edição e de carga de definições de classes de entidades espaciais num programa de computador e, ainda, agregar funcionalidades mais específicas, como submeter esse *corpus* a cálculos de similaridade semântica entre essas classes de entidades.

O que é de unânime consenso entre os cientistas da computação e lingüistas computacionais é que, na ausência de uma especialidade daquela engenharia para resolver esses problemas, as ferramentas de modelagem e de programação da OO são as mais adequadas por enquanto.

3.3. Síntese da revisão de literatura

Os seguintes tópicos revisados trazem elementos que serão utilizados no refinamento do objetivo geral desta pesquisa, bem como na formulação da hipótese de pesquisa:

- Na modelagem de um SIG de utilidade geral, é igualmente importante atentar para o levantamento de requisitos de ordem cognitiva além do de ordem formal (geométrica);

- Um estudo na área da Ciência da Informação que trate da produção de *software*, necessariamente, exige a construção de um protótipo que antecipe o desenvolvimento de um futuro sistema;
- A linguagem de modelagem e a de programação a serem utilizadas na construção de um protótipo que se destina a representar um problema complexo como o da avaliação da SS, já devem espelhar o tipo de uso que se fará do sistema futuro que irá incorporá-lo. Se este trata de modelagem da informação num nível bem próximo da maneira humana de raciocinar, envolvendo percepção do MR, indução de propriedades do fenômeno e abstração na solução, é recomendado o uso de linguagens que incorporem os fundamentos da OO, mais adequadas à implementação de problemas de IA, por enquanto;
- As especificações deste protótipo devem retratar o princípio da parcimônia científica ou da *Navalha de Ockham*:

"A hipótese mais provável é aquela mais simples e consistente com todas as observações."

A seguir, será comprovado por que a opção de RODRÍGUEZ (2000)[207] em trabalhar num subconjunto da LN, como a LP do domínio da CIGeo, deve ser calcada na Filosofia da Linguagem e na Lógica para ser plausível. É desta LP que foram extraídos o *corpus* sobre o qual a autora aplicou o seu MSS e, de igual forma, de onde também foi extraído parte do *corpus* sobre o qual o PRONTO© foi construído.

Adaptando os conceitos de denotação e de conotação de PINTO (1977)[208] ao fenômeno que foi observado e sobre o qual se fizeram experimentos nesta tese, por uma perspectiva exploratória, pode-se formular o seguinte enunciado:

"Se a classe de entidades espaciais 'X' possui o conjunto 'Y' de feições distintivas (características ou *distinguishing features*), então o termo 'X', que a ela se refere, é denotativo."

Analisando o enunciado sob a luz da Lógica e da IA (ciência cognitiva), é possível tirar algumas conclusões.

A primeira conclusão é que, à luz da Lógica, o enunciado pode transformar-se num argumento muito comum, empregado na criação de inferências (passos de uma demonstração lógica), que é o "*modus ponens*" (afirmação do antecedente), bastando apenas o seguinte: considerá-lo como premissa, afirmar o seu antecedente e concluir com o seu conseqüente, desta maneira:

p $\Rightarrow$ **q:** "Se a classe de entidades espaciais 'X' ..., então o termo 'X' ... é denotativo." **::=** implicação

---

[207] E, por extensão, a opção também assumida nesta tese.
[208] V. p. 159.

p: "A classe de entidades espaciais 'X' possui o conjunto 'Y' de características" **::=** antecedente

q: "O termo 'X', que a ela se refere, é denotativo." **::=** conseqüente

Apesar de a Lógica não se interessar pela verdade ou falsidade das evidências das premissas, este argumento será válido se as premissas forem verdadeiras e se derem sustentação à conclusão, que também deve ser verdadeira, de acordo com a fórmula do *modus ponens*.

É bom lembrar que um argumento é um conjunto de enunciados (premissas e conclusão) relacionados entre si; que uma premissa é um enunciado que contém evidência e que uma premissa verdadeira constitui uma evidência ou um enunciado legítimo de um fato, que é do interesse da ciência, ou seja, a ciência se preocupa com o conteúdo dos enunciados e a Lógica, só com a forma deles; e que enunciados são verdadeiros ou falsos, mas argumentos são válidos ou não-válidos, logicamente.

A segunda conclusão prende-se ao fato de que é preciso investigar cientificamente o conteúdo dos enunciados, o que só se pode realizar pela metodologia do ramo da ciência em questão.

Sendo assim, o argumento que rege esta pesquisa é logicamente plausível e, no aspecto científico, este argumento necessita de apoio metodológico para comprovar a hipótese de pesquisa. A metodologia revelará a verdade das premissas, i.e., por intermédio dela é que se demonstrará a existência de evidência nas premissas, o que só a avaliação da SS proporcionará, ao ajudar a preparar o caminho para a tão decantada e perseguida *Teoria Semântica*.

4. FORMULAÇÃO DA HIPÓTESE DE PESQUISA

"Aprender a desconfiar de si, e notadamente quando se está mais seguro de ter razão, suspeitar de toda a afirmação não acompanhada de prova, inclinar-se diante de toda verdade bem demonstrada, mesmo que ela perturbe, desiluda, negue; eis aí a boa lição que não é apenas de vigilância e disciplina intelectual, mas também de higiene moral." (Jean Rostand, biólogo francês)

A opção por enunciar a hipótese de pesquisa após a revisão de literatura, teve o intuito de refinar o seu enunciado, limitando-o aos elementos mais susceptíveis de integrar o fenômeno da SS e para orientar a procura da solução para o problema levantado, o que servirá de subsídio para pormenorizar o objetivo geral.

Teve-se o cuidado de empregar nos enunciados das hipóteses termos já definidos em itens anteriores, considerando-se ainda as restrições do subitem 4.3.

A fim de facilitar a visualização da relação lógica que deve existir entre o problema geral e sua hipótese (pressuposto) de pesquisa ou de trabalho, volta-se a apresentar o enunciado do primeiro:

"Como avaliar a similaridade semântica entre objetos espaciais representáveis numa base de dados?"

Um retorno ao esquema de análise lógica (*rationale*), exposta na Figura 2.3 (subitem 2.3.1) será de muito bom alvitre neste ponto do texto.

4.1. Hipótese alternativa

"Nada é bom ou mau, a não ser por comparação."
[Thomas Fuller (*apud* WONNACOTT (1980, p. 247)]

O enunciado da hipótese alternativa de pesquisa, que será operacionalizada no subitem 5.3, é o seguinte:

H - 1:

"A utilização de um protótipo para avaliar a similaridade semântica entre objetos espaciais, ***representáveis*** *numa base de dados cartográficos, é um método automatizado e compatível com o critério humano de julgamento de similaridade para distinguir entidades geográficas no mundo-real."*

4.2. Hipótese nula

A manutenção do enunciado da hipótese nula se justifica pela aplicação da metodologia tradicional de uma pesquisa experimental, que é buscar uma prova estatística boa o suficiente, i.e., de pequena probabilidade para rejeitá-la, se for verdadeira, ou de pequena probabilidade de aceitá-la, se for falsa.

Esta metodologia é mais eficaz em termos estatísticos, porque a hipótese nula é mais fácil de ser transformada em termos quantitativos, conforme o artifício estatístico de que é mais trivial provar que um argumento (conjunto de enunciados) não é válido do que a sua validade. Por mais que se esmere na juntada de evidências para provar que os enunciados de um argumento estão carregados de evidência, basta que um deles venha a ser considerado como falso, pela simples apresentação de uma inconsistência numa das evidências, e todo o argumento, inicialmente considerado como sólido, poderá debilitar-se ou até mesmo invalidar-se, ou seja, a sua conclusão poderá ficar comprometida[209].

O enunciado da correspondente hipótese nula é o seguinte:

H - 0:

"Não existe uma correlação significativa entre o resultado alcançado por um protótipo para avaliar a similaridade semântica entre objetos espaciais, ***representáveis*** *numa base de dados cartográficos, e o resultado apurado pelo critério humano de julgamento de similaridade para distinguir entidades geográficas no mundo-real."*

Quando for o momento de operacionalizar o enunciado da hipótese nula, o seu enunciado será transformado em duas hipóteses estatísticas, ligadas à fase de avaliação do PRONTO© com relação às respostas dos questionários distribuídos aos indivíduos do segundo campo de observações, seguindo, em linhas gerais, a metodologia estabelecida por RODRÍGUEZ (2000).

4.3. Considerações de restrição à hipótese

A inclusão do termo "base de dados" na formulação dos enunciados do problema, do objetivo e da hipótese de pesquisa precisa de maiores explicações.

Originalmente, nos enunciados dos tópicos acima, o verbo iterativo "representar" foi empregado em sua forma nominal do particípio ["representado(s)"], para se referir aos obje-

[209] No mundo judicial, isto é bem comum: não importa o esforço da defesa em juntar provas para defender o seu ponto de vista; se o promotor derrubar uma das provas, todo o esforço da defesa poderá ser desqualificado.

tos espaciais, presumivelmente armazenados e organizados numa estrutura sistematizada de arquivos, como é o caso de uma base de dados.

Tal cenário, todavia, não se evidenciou, visto que, como será explicado, o custo da pesquisa tornar-se-ia proibitivo em termos materiais, tecnológicos e mesmo legais para sustentar o rigor do emprego do verbo *representar* na sua forma nominal do particípio ("objetos *representados* numa base de dados").

Como ensina ALMEIDA (1994, p. 434), os verbos terminados em "do(a)" são formas nominais adjetivadas, que exprimem uma circunstância de tempo definida ou uma etapa de ação concluída (fim de um estado transitório).

Dessa forma, não seria coerente manter os enunciados desses tópicos com esta forma nominal do verbo "representar". Os objetivos específicos da pesquisa ficariam comprometidos, já que não se trabalhou com dados estruturados em SGBDs.

Mas o mesmo ALMEIDA (1994, p. 393) ajudou a solucionar morfologicamente o problema de se obter um enunciado coerente. O autor também ensinou que a classe de adjetivos terminada em ***ável*** indica aptidão, estado e *possibilidade de praticar ou de receber uma ação*, o que se ajusta perfeitamente ao adjetivo "representáveis"[210], que foi a forma escolhida para os enunciados dos tópicos em pauta. Esta forma traduz uma possibilidade e não um fato consumado de representação, o que é mais palpável e exeqüível diante dos meios disponíveis.

Mas esta solução morfológica para adaptar os enunciados desdobrou-se num outro problema: Há fundamentação na Ciência da Computação para se estudar um fenômeno formalizável em *bits* e *bytes* sem utilizar as tradicionais soluções de processamento de dados (**PD**), fortemente dependentes de um sistema que perenize essas cadeias de zeros e uns em arquivos? Esta pergunta é muito pertinente, porque *objetos representáveis*, i.e., que poderão ser *objetos representados*, pelo que se levantou na revisão de literatura até meados de agosto de 2002, não poderiam ser tratados pelos procedimentos ortodoxos de PD, especialmente os SGBDs de natureza relacional, os mais difundidos.

Esta restrição parecia transforma-se numa dificuldade intransponível, um verdadeiro *nó górdio* de uma pesquisa que também é de cunho experimental.

Este *nó górdio* só pôde ser desatado em 15 de agosto de 2002, por acaso, ao se aceitar um convite de TABALIPA (2002)[211] para uma palestra do DFJUG[212].

---

[210] Em virtude dessa minúcia discutida, o vocábulo foi sempre realçado nesses enunciados.

[211] Também: M.C. e analista *sênior* da ANVISA/Min. Saúde (www.anvisa.gov.br), Coordenador-Chefe do Curso de Sistemas de Informação da Faculdade Alvorada

[212] Grupo de Usuários de Java do DF - www.dfjug.org .

TABALIPA (2002) garantiu ser possível trabalhar num nível de abstração bem alto, como no de modelagem conceitual, em que os objetos não necessitam ser transformados em estruturas de dados convencionais de níveis mais baixos da cadeia de abstração de desenvolvimento de sistemas.

Nessa palestra, o pesquisador Klaus Wuestefeld, pioneiro da *XP* (***EX****treme* ***P****rogramming*) no Brasil, demonstrou que é possível implementar toda a funcionalidade de um SGBD em nível conceitual, no interior dos objetos da linguagem *Java*™. Esse pesquisador brasileiro é o criador do PREVAYLER™ – www.prevayler.org), um sistema que explora o conceito de prevalência (hegemonia) de objetos de *Java*™ (DFJUG, 2002).

BRITO (2003) procurou descrever brevemente o conceito de *prevalência*. O autor começou a descrição, explicando que a maior parte das aplicações na área de PD e de sistemas de informação necessita de alguma forma de *persistência*, ou seja, de conseguir fazer com que os dados "sobrevivam" na memória RAM do sistema computacional durante o momento de execução de uma certa aplicação. Essa preservação (armazenamento) dos dados é que garante a progressão da execução do programa, numa seqüência controlada pelo sistema operacional.

Os mecanismos mais comuns de preservação dos dados são os sistemas de arquivos, as bases de dados, entre outros. O maior problema apresentado por esses meios é que eles não reproduzem o poder de abstração da OO. Para que esses objetos da RAM sejam armazenados num meio perene de memória (disco-rígido), eles devem passar por processos complicados e demorados (em termos computacionais) de descaracterização, como p.ex., os *motores de persistência*. Esse processo baseia-se num mapeamento especificado num arquivo do tipo XML™[213], para transformar classes da OO em tabelas de BDs, ou propriedades das classes da OO em campos das BDs (BRITO, 2003).

Tanto para BRITO (2003) como para WUESTEFELD (2003a), esses processos de mapeamento "desmontam" o encapsulamento dos objetos, que é um conceito fundamental da OO. Além disso, consomem preciosos recursos computacionais e constituem-se em prováveis fontes de erros para o sistema de informação em desenvolvimento.

Daí surgiu a inquietação necessária para que WUESTEFELD (2003a) criasse o conceito de prevalência, que se baseia num princípio muito simples, mas surpreendentemente muito eficaz: armazenar os objetos na volátil memória RAM em vez de numa BD e, para compensar essa desvantagem da volatilidade desse tipo de memória, somente em momentos

[213] *EXtensible Markup Language*, metalinguagem e subconjunto da SGML (*Standard Generalized Markup Language*), ambas criadas pelo Consórcio W3C (WWW *Consortium*), especificamente para a representação de dados.

críticos da execução de um programa é que os dados são descarregados num arquivo de segurança, denominado de *arquivo log*[214]. Dessa forma, a integridade do sistema é mantida sem a perda do desempenho que normalmente acompanharia essa fase da execução de um programa carregado na RAM.

É interessante citar o resumo que WUESTEFELD (2003a) enviou para os organizadores (DFJUG, 2002) da palestra mencionada *ut retro*, que traduz o estado de espírito que estimulou o pesquisador em sua descoberta:

"Eu não agüentava mais a lentidão dos sistemas gerenciadores de bancos de dados.

Eu não agüentava mais ficar instalando o banco de dados no escritório, em casa, para cada cliente e em todo lugar onde eu queria programar ou demonstrar os meus sistemas. Frustrava-me não usar todo o poder de objetos de verdade, como eu bem entendesse. Eu detestava a poluição do pré e do pós-processamento que os bancos de dados OO causavam às minhas classes. Eu achava ridículo ter que distorcer as minhas classes e ficar criando mapeamentos com tabelas em um banco de dados relacional. Acima de tudo, revoltava-me ver as pessoas reclamando que as promessas da OO não se cumpriam."

O PREVAYLER™ é realmente revolucionário no campo da Ciência da Computação, em particular na área de desenvolvimento de sistemas, porque acena para o ocaso de SGBSs consagrados como o ORACLE™ e o MySQL™[215], em alguns tipos de problemas (será que, algum dia, para todos os problemas ligados a SGBDs?). O produto em pauta chega a ser cerca de dez mil vezes mais rápido do que o primeiro e três mil vezes mais rápido do que o segundo. E o que é mais impressionante: é de concepção simples e é gratuito.

O PREVAYLER™ é o resultado da implementação do conceito de prevalência. A maneira de se trabalhar com esse produto é muito trivial: criam-se os objetos que representam o modelo de dados do usuário, uma classe que se destina a ser o *contêiner* dos resultados e as operações que serão aplicadas aos dados.

As características do PREVAYLER™, enumeradas por WUESTEFELD (2003a), demonstram por que é possível trabalhar com classes de objetos em vez de trabalhar com bancos de dados tradicionais:

- O arquivo *log* é a solução mais simples para evitar a perda de dados durante a execução de um sistema de informações que dispensa os recursos de um SGBD. Essa operação é como uma sucessão de fotos (*snapshots*) que "congelam" um determinado estado da e-

[214] São arquivos que registram os sucessivos estados de execução de um programa.
[215] São produtos da *Oracle, Corp.*. e da *MySQL, Inc.*, respectivamente.

xecução, garantindo a sua continuidade, mesmo em situações críticas (queda de energia, p.ex.). Tal operação é similar ao estado de hibernação de um *notebook*.

- A memória RAM, tanto em qualidade como em quantidade, é o ponto fulcral do PREVAYLER™. Já há relatos de avanços consideráveis na tecnologia de fabricação de RAM (RAM holográfica, RAM magnética e RAM polimerizada). Algumas dessas memórias já alcançam taxas de transferência de 1 GB/s e têm capacidade de armazenamento de até 1 TB, amenizando ou eliminando as vulnerabilidades do produto nesse aspecto (WUESTEFELD, 2003b).
- O PREVAYLER™ é um produto que implementa a prevalência de objetos na linguagem *Java*™ e foi classificado quanto à sua distribuição como um produto *freeware*, i.e., não custa nada, porque foi licenciado sob os auspícios da *Free Software Foundation* pela LGPL (*Lesser General Public License*). A sua licença de código aberto propicia maior participação da comunidade de usuários, que examinam o produto, melhoram-no em suas deficiências e ampliam as suas capacidades de aplicação.
- A prevalência de objetos é muito mais rápida que um SGBD tradicional, porque opera sobre objetos em sua forma nativa. É desnecessário gerar e efetuar comandos numa linguagem de manipulação de dados[216]. Não há restrição quanto às linguagens para escrever os algoritmos para as estruturas de dados. É difícil ser mais rápido do que isso.
- Há compatibilidade total dos objetos gerados e operados pelo PREVAYLER™ com aplicativos-clientes, executados noutros sistemas[217] comuns do mercado.
- O PREVAYLER™ também permite a implantação dos seus objetos num ambiente de rede mundial (www).
- O PREVAYLER™ faz cumprir a missão original atribuída à OO, ao desviar o foco da geração de código num servidor complexo de dados para a geração de código num servidor simples de objetos, mesmo que o usuário arrume as mais diversas desculpas para continuar ligado (emocionalmente) ao seu "velho banco de dados". Como disse WUESTEFELD (2003a): "Pelo menos, agora, o usuário tem uma alternativa".

Para encerrar essa parte sobre os depoimentos e descobertas recentes que dão suporte à manipulação experimental desta pesquisa num "nível alto de abstração", ainda é o caso citar-se o trabalho de GONÇALVES (2001), que também explorou conceitos do *Geoprocessamento*, apesar de ter construído um protótipo que opera com elementos estruturais das imagens e não com termos.

---

[216] V. quarto nível do modelo de SETZER (1989), no subitem 1.5.3.1.
[217] CORBA©, p.ex.

GONÇALVES (2001, p.21), explicitamente, declarou que seu protótipo gerou conteúdos de informação (e não de dados). A implementação do seu protótipo também seguiu alguns princípios de PLN[218]. Nesse aspecto, seu objetivo foi o de propiciar que o seu protótipo fosse capaz de trabalhar em "níveis maiores de abstração" (*descritores contextuais* representados numa metalinguagem), para criar diferentes relações entre as classes de entidades do modelo conceitual, concebido para a recuperação de informação temática contida nessa forma peculiar de documento, que é a imagem, não diferindo em quase nada do propósito que anteriormente se procurou justificar para o uso do adjetivo "representáveis" em vez da forma verbo-nominal (particípio) "representadas".

[218] Além da Teoria dos Grafos e da Teoria dos Autômatos

5. PROBLEMA, OBJETIVOS E HIPÓTESES - ASPECTOS ESPECÍFICOS

"Não diga – Encontrei a verdade! -, mas antes – Encontrei uma verdade!"
[Kahlil Gibran (*apud* WONNACOTT (1980, p. 215)]

5.1. Delimitação do problema de pesquisa

Em relação ao conceito de *base de dados*, a real dimensão do que se desejou exprimir nos enunciados dos tópicos genéricos da pesquisa (problema e objetivo gerais, assim como a hipótese de pesquisa) serviu de preâmbulo para este capítulo.

Apenas para fixar melhor esse conceito no âmbito da CI, visto que na Ciência da Computação há poucas divergências a respeito, é instrutivo estender um pouco mais as considerações do capítulo precedente, agora por um prisma mais didático.

Como já antecipado, ambos os protótipos não são aplicados sobre *bases de dados*, no sentido de rigor que o termo denota. Para deixar isso bem explícito, em vez de se usar a expressão "representados" para qualificar os objetos espaciais, foi utilizada, propositadamente, o termo "representáveis", da formulação do problema geral até a formulação da hipótese nula.

Bases de dados[219], a rigor, devem ser estruturadas para permitir que o SGBD represente a realidade, composta de entidades, num modelo que contenha objetos e suas relações entre si. Esses objetos relacionados poderão estar disponíveis para diversas rotinas de SGBD, que têm como principal objetivo cruzar dados para derivar informações.

O emprego do adjetivo "representáveis" é muito mais adequado nesses tópicos genéricos da pesquisa, visto que traduz a possibilidade e não o fato consumado da reprodução dos dados, aos quais ele se refere, num meio perene de memória, susceptível de ser utilizado pelas mais diversas aplicações computacionais.

A alusão ao termo *base de dados* não expressaria o domínio real do problema, se não fossem dadas tais explicações. Nenhum dos dois protótipos trabalhou sobre bases de dados e nem tampouco interessa, nesta pesquisa, estudar as classes de entidades espaciais e suas relações nesse nível de abstração, ou seja, no quarto nível de SETZER (1999, p.5) ou *nível operacional* (dados a serem carregados num sistema computacional).

---

[219] Para YONG (1983), banco de dados = {base de dados + SGBD}, admitindo a confusão entre base e banco de dados por simplificação terminológica.

A alusão ao termo *base de dados* teve o propósito de dar prosseguimento à concitação de trabalho futuro de RODRÍGUEZ (2000, p. 138). No entanto, as presentes observações são necessárias para deixar mais claras as nuanças conceituais que envolvem as estruturas de dados denotadas por esse termo e as estruturas de dados que foram realmente utilizadas no PROFAX e no PRONTO©, atinentes a seqüências, condições e repetições que mantêm o controle sobre um conjunto de *arquivos orientados para programa* ou *coleção de dados orientada para programa* (CDOP).

Uma CDOP é mais limitada em termos de integridade do que uma base de dados. YONG (1983, p.18-31) estabelece características distintivas entre os dois tipos de sistemas que operam os dados sob seus domínios. Sobre uma base de dados opera um SGBD, como já explicado. Mas no caso de uma CDOP, o que existe é um *sistema de processamento de dados orientado para programa* (SPDOP), ou o equivalente ao módulo dos protótipos[220] encarregado de gerenciar dados (ler, excluir, carregar, etc.).

Tabela 5.1: Diferenças entre um SPDOP e um banco de dados.

	SPDOP	Banco de Dados
CARACTERÍSTICAS	**1.** Arquivos criados de acordo com as necessidades do sistema.	**1.** O banco de dados deve focar as entidades do mundo e representá-las num modelo conceitual capaz de traduzir as propriedades e as relações entre estas entidades, independente de necessidades contingenciais.
	2. As coleções de dados são projetadas para atender às necessidades do sistema.	**2.** Os dados estruturados num banco de dados devem possuir a devida perenidade para atender a diversos tipos de aplicações exógenas ao banco.
	3. O centro de gravidade do sistema são os programas.	**3.** O centro de gravidade do banco são os programas que constituem o SGBD.
	4. Há normalmente uma grande utilização de classificações.	**4.** O banco de dados minimiza a utilização de classificações.
	5. O acesso aos dados do arquivo é feito diretamente pelo programa de aplicação.	**5.** O acesso aos dados do banco é feito pela camada de *software* representada pelo SGBD, que fica entre a aplicação do usuário e a base de dados.

No presente caso, como já parcialmente explicado no final do subitem 1.2.3, os dois protótipos não foram construídos com mecanismos sofisticados de inferência para rastrear ontologias, mas as suas estruturas de dados, combinadas com a taxinomia e o *corpus* construído, de certa forma, aproximam essas ferramentas de validação de hipóteses de um AI.

[220] Lembrar que os protótipos já foram classificados como um ARS (agente reflexo simples), algo mais que um SPDOP.

A Tabela 5.1 identifica as principais características de um SPDOP e de um banco de dados, para que se tenha uma idéia da correspondente relação entre uma CDOP (de uso no PROFAX) e uma base de dados (de uso imaginado para extensões do PRONTO©).

Como se verá na metodologia, as estruturas dos protótipos superam em boa parte as limitações de um SPDOP no tratamento de dados. A implementação do protótipo por uma LTP como o *Java*™, incorporando muitos conceitos da OO e executando a fórmula matemática da relação de SS aplicada a uma taxinomia preparada com base na modelagem conceitual de uma verdadeira base de dados, como a carta Faxinal, são três requisitos que, acredita-se neste ponto da pesquisa, podem substituir uma linguagem de representação do conhecimento incorporada num ABM, como poderá ser o futuro do PRONTO©. Somente a realização dos testes e a análise dos resultados poderão confirmar esta convicção do pesquisador em barganhar com simplificações para obter resultados satisfatórios.

Para o leitor que não é ou não está afeito ao campo *geocientífico*, é instrutivo saber que esse traço da pesquisa, baseado em simplificações e aproximações, vem plasmado em todas as cadeiras ministradas em cursos como Cartografia, Astronomia, Geofísica, Geodésia, Topografia e outros. Determinar posições no espaço terrestre (ou extraterrestre) sempre implicou converter formas matematicamente intratáveis, como p.ex., um *geóide*[221], que representa a Terra, em superfícies e sólidos sobre os quais possam ser aplicadas formulações matemáticas conhecidas, como uma esfera para representar uma aproximação de um *geóide* (da Terra real).

Portanto, voltando à questão discutida no início deste subitem, que é a de restringir a generalidade do primeiro objetivo de pesquisa orlado no subitem 2.3 e repetido na hipótese (o termo *base de dados*), este pesquisador pretende aduzir para a metodologia elementos que poderão sustentar este seu juízo de simplificação sem comprometer o objetivo colimado.

Não serão explicitados os problemas específicos de pesquisa em favor da simplicidade, já que eles podem ser obtidos da transposição de cada um dos três objetivos específicos a seguir (na forma afirmativa) para a forma interrogativa.

5.2. Estabelecimento dos objetivos específicos de pesquisa

A fim de facilitar a visualização da relação de continuidade entre o objetivo geral e os específicos, volta-se a apresentar o primeiro:

[221] É comum chamar esta figura de um pêra, em cursos de 1° grau.

"Avaliar a similaridade semântica entre objetos espaciais representáveis numa base de dados."

Seguem-se os **objetivos específicos (OE)** de pesquisa:

- **OE-1: determinar** os indicadores para o julgamento humano da similaridade semântica entre classes de entidades espaciais;
- **OE-2: determinar** os indicadores para a avaliação computacional da similaridade semântica entre classes de entidades espaciais, representáveis numa base de dados;
- **OE-3: comparar** os resultados do julgamento humano e a avaliação computacional da similaridade semântica entre classes de entidades espaciais, representáveis numa base de dados.

5.3. Formulação das hipóteses estatísticas

Antes da enunciação das duas hipóteses que reduzirão as informações dos elementos do problema em termos numéricos, ou que operacionalizarão a hipótese de pesquisa, é instrutivo saber, em linhas gerais, como a autora da tese da qual esta se originou tratou desse tópico da pesquisa.

RODRÍGUEZ (2000) desenvolveu o MSS em duas fases: 1ª) Estabeleceu uma função de cálculo de distância semântica; 2ª) Combinou a primeira função com uma outra: a de verificação de feições distintivas (visão da Psicologia Cognitiva). Essa função combinada propiciou ao MSS resolver o relacionamento semântico entre as classes de entidades espaciais (relações de gênero-espécie e meronímicas) e discriminar cada instância de classe pela verificação de similaridade (ou diferença) entre as funções distintivas (partes, funções e atributos) de cada classe. Numa ulterior versão do MSS, foi incorporada a capacidade de inferência contextual, pela introdução de axiomas de explicitação do contexto na ontologia que a autora construiu.

A construção do PROFAX representou a primeira fase:experimental da pesquisa, para se verificar o comportamento da função de distância semântica numa hierarquia. O projeto de pesquisa não previu agregar a esse primeiro protótipo a capacidade de distinguir instâncias de uma classe, p.ex: ele ainda não conseguia avaliar semanticamente um rio e uma corredeira, que pertencem à mesma classe hidrográfica. O contexto, como já declarado, não foi incluído explicitamente nesse protótipo. Metaforicamente, o que se desejava nessa fase do trabalho era "pôr no mesmo guarda-chuva" as instâncias, segundo seus elementos básí-

cos de significado; em seguida, "pulverizá-las" (individualizá-las), dentro de cada "guarda-chuva", de acordo com suas propriedades ou feições distintivas.

Na sua análise de resultados, RODRÍGUEZ (2000, p.98) apresentou as pontuações em graus de correlação, que variaram de [0, 1], entre três classes de modelos de avaliação de SS, que servirão como referencial para esta pesquisa. Os modelos geométricos alcançaram uma correlação média de 0,60. Os modelos baseados na Teoria da Informação alcançaram a média de 0,79 e os que usaram funções baseadas na verificação de funções distintivas, 0,83.

A autora não entrou em minúcias sobre como foram construídos os modelos geométricos testados, mas já foram separados três trabalhos que utilizaram modelos geométricos da forma pela qual se pretendeu resolver esse problema da diferenciação de classes. Esses trabalhos são os de WONG (2000), GANESAN (2002) e SANTOS (2002). Na órbita desses três, foram separados muitos outros que ainda não foram examinados detidamente e que podem ser objeto de investigação futura. Os três trabalhos desses autores já foram suficientes para compor o referencial teórico-experimental da primeira fase da pesquisa, em que o PROFAX foi bem-sucedido.

Para a segunda fase, de cujo projeto de pesquisa emergiu o PRONTO©, para operar em conjunto com o questionário aplicado a profissionais do *geoprocessamento*, a idéia que já se tinha sobre essa função diferenciadora de instâncias de classes era a de utilizar o potencial da OO, de tal forma que os *elementos de definição*[222] dos termos referentes às entidades espaciais pudessem passar como atributos de objetos do MC da folha Faxinal.

Além das injunções e recomendações até aqui citadas, as hipóteses estatísticas deverão levar em conta o que foi grifado no último parágrafo*: elementos de definição*. A metodologia estenderá o assunto, mas a formulação dessas hipóteses deverá considerar a montagem de definições precisas das classes de entidades espaciais. RODRÍGUEZ (2000) reputou essa fase como um importante passo na construção do seu MSS. Definições de termos precisas implicam cálculo de SS precisa. Portanto, surgiu mais essa tarefa na pormenorização do projeto de pesquisa: construir definições e montar a árvore *n-ária* (taxinomia*)*.

Diante de todos os apontamentos anteriormente delineados, pode-se esboçar os tópicos frasais que orientarão a elaboração dos enunciados das hipóteses estatísticas:

- *A primeira hipótese deve fixar-se na comparação das respostas dos indivíduos;*

---

[222] Relações hierárquicas e feições distintivas de cada classe de entidades espaciais.

- *A segunda hipótese deve fixar-se na comparação entre as respostas dos indivíduos com os resultados do PRONTO©.*

5.3.1. Variáveis de pesquisa (nível geral) e enunciado das hipóteses estatísticas

No subitem anterior foi citada a intenção de levantar as variáveis gerais, uma vez que a demonstração do funcionamento do PROFAX, com a execução da função de SS para relações de gênero-espécie, é um indicativo de que já se trabalha com uma variável empírica dependente (tipo: razão[223]): a ***similaridade semântica***, bastando consultar a Equação 3.1. O co-seno do ângulo entre os dois vetores da Figura 3.12 é a representação da similaridade semântica.

Aí está uma das utilidades do primeiro protótipo, ao permitir que o pesquisador visualizasse os contornos da variável dependente, à proporção que montava o exemplo de treinamento para este caso de avaliação de SS, bem limitado em seus resultados.

Da Equação 3.1, ainda se pode identificar variáveis empíricas independentes no segundo membro da equação, que permanecerão no modelo matemático (MM) do PRONTO©.

Ao analisar o numerador do segundo membro da equação em tela, percebe-se um produto interno entre dois vetores. O denominador é o produto das normas de ambos os vetores. Esses vetores, por enquanto, são formados por coleções de instâncias de classe (conforme já foi dito, o PROFAX não mede a SS de uma instância em relação a outra instância). Para identificar as variáveis independentes nesse MM, basta verificar como foi calculada a SS entre coleções de instâncias segundo GANESAN (2002). As coleções de instâncias é que são os vetores e os componentes desses vetores são as instâncias de classes (rio, lago, estrada, etc.), que são quantificadas pelas posições que ocupam na árvore *n-ária*, com base no que vários dos autores pesquisados chamam de ***lub*** (*least upper bound*) ou o ancestral de ordem mais baixa (inferior) da instância, na taxinomia (árvore). É dele que se contam as braçadas (arcos) até o nó-raiz[224] (ou patriarca) da árvore. É o ***lub***, por conseguinte, uma das variáveis independentes do PROFAX. A **posição das instâncias na árvore** pode ser considerada como variável intermediária.

Por que, no parágrafo anterior, foi dito que o ***lub*** é uma das variáveis independentes? É porque as hipóteses estatísticas, nessa altura da pesquisa, ainda não tinham sido formuladas e poderiam surgir outras, como já se supunha, diante da limitação da fórmula do co-seno para calcular a SS entre instâncias das classes. Pelo que se presumia, quando fosse

---

[223] Classificação de RICHARDSON (1999, p.126-129).

[224] Lembrar que esta é uma árvore invertida, segundo a metodologia da Ciência da Computação.

necessário diferenciar cada uma das instâncias por suas propriedades (*feições distintivas* ou *fds*), uma ou mais variáveis independentes surgiriam.

Apenas para se ter uma visão geral do levantamento de variáveis para o PRONTO©, a fórmulas (*rationale*) abaixo e as descrições correspondentes são apresentadas:

- **Z = *f* (X, Y):** capacidade de avaliação da SS do protótipo
- **X:** *internodalidade* [enfoque de rede de RIPS (1973)]
- **Y:** grau de superposição (*feature matching*) de feições[225]
- **W = *f* (Z, Resp):** correlação entre Z (SS) e as respostas (Resp) ao questionário

O exercício que o PROFAX permitiu executar (subitem 6.3.1.1) é que proporcionou as condições necessárias de esboçar a formulação esquemática anterior e a visualização das variáveis gerais e intermediárias de pesquisa, passo fundamental para a orladura das hipóteses estatísticas e a posterior derivação das variáveis empíricas.

De acordo com o *rationale* anterior, já é possível antecipar a comparação entre os resultados obtidos dos dois campos de observações (indivíduos e PRONTO©). Esta comparação sobrevém do cálculo da correlação explicitada na última fórmula esquemática.

No caso do campo de observações dos indivíduos, o instrumento de coleta de dados que será empregado é o tradicional questionário, cuja elaboração obedeceu aos seguintes requisitos de ordem geral:

- Perguntas fechadas de alternativas (gradação de 1 a 10), exaustivas e excludentes;
- Público especializado (engenheiros e técnicos de *geoprocessamento*);
- Definições preliminares sobre as classes de entidades espaciais;
- Subdivisão[226] do questionário: 1) Relações gênero-espécie; 2) Relações todo-parte (testam a assimetria);
- Classes de entidades do PRONTO devem constar do questionário;
- Perguntas simples, sem considerar o contexto.

Os requisitos do questionário – um dos instrumentos de validação da hipótese de pesquisa – também servem como parâmetros de orientação para a formulação das hipóteses estatísticas e para o levantamento das variáveis empíricas.

Para facilitar a visualização da transformação da hipótese de pesquisa nas duas hipóteses estatísticas que aferirão o PRONTO©, volta-se a apresentar a hipótese de pesquisa:

H-1: "*A implementação de um protótipo para avaliar a similaridade semântica entre objetos espaciais, representáveis numa base de dados cartográficos, é um método automati-*

[225] Enfoque da Teoria dos Conjuntos.

zado compatível com o critério humano de julgamento de similaridade para distinguir entidades geográficas no mundo-real."

Os enunciados das hipóteses estatísticas, i.e., as que operacionalizarão a hipótese nula de pesquisa vêm a seguir:

He1: "As respostas dos operadores humanos estão relacionadas."

He2: "Os julgamentos de SS dos indivíduos e os resultados da avaliação de SS do PRONTO© estão correlacionados."

Nessa fase da pesquisa, por se tratar de uma continuação do trabalho de RODRÍGUEZ (2000), supunha-se que o modelo de cálculo estatístico a ser aplicado na fase seguinte fosse o mesmo (o que se confirmou). Por conseguinte, foram utilizados dois indicadores não-paramétricos para medir o nível de correlação entre as respostas dos indivíduos entre si e entre as respostas dos indivíduos e os resultados correspondentes, atingidos pelo PRONTO©.

No primeiro caso anteriormente mencionado (He1), foi utilizada a fórmula de cálculo do coeficiente não-paramétrico de concordância *W de Kendall* entre as respostas (variável dependente desse modelo estatístico). As variáveis independentes desse modelo são o número de indivíduos que responderam às perguntas e a ordenação (postos) de cada resposta. A primeira hipótese estatística serve também para eliminar *outliers* (observações afastadas do corpo da amostra).

No caso da segunda hipótese estatística (He2), foi utilizada a fórmula de cálculo do coeficiente de correlação R_s *de Spearman* (variável dependente desse modelo estatístico). As variáveis independentes desse modelo são o número de indivíduos que responderam e a diferença entre o posto obtido numa avaliação do PRONTO© e o posto da resposta correspondente do indivíduo.

As Equações 6.4 a 6.7 do subitem 6.3 explicitam esses modelos estatísticos.

5.3.2. Variáveis empíricas (nível específico)

Encerrando o capítulo, com base no referencial teórico que dá sustentação a esta pesquisa, já é possível extrair as variáveis empíricas que comporão o MM (Equações 6.1 a 6.3) de avaliação de SS do PRONTO© e o modelo estatístico. Este último constitui a base de desenvolvimento da metodologia, que se segue no Capítulo 6.

Admitindo-se a formulação esquemática ***z = f (x, y)***:

---

[226] Ambas as partes tendo por sujeito (variante) a mesma classe.

- $z \Rightarrow$ ***SS***; variável dependente e de razão, estabelecida para avaliar a similaridade semântica entre classes de entidades espaciais representadas num subconjunto do MC da folha Faxinal e que varia no intervalo [0,1] do conjunto $\Re$. As Equações 6.1 e 6.2 formalizam o seu cálculo;
- $x \Rightarrow$ ***lub;*** variável independente e intervalar do conjunto $\aleph$, estabelecida para determinar a distância do ancestral imediatamente comum entre as classes de entidades sob avaliação de SS e o nó-raiz da hierarquia. A Equação 6.3 formaliza o seu cálculo;
- $y \Rightarrow$ ***contagem*** *(freqüência) das feições comuns segundo partes (y_1), funções (y_2) e atributos (y_3);* variável independente e intervalar, estabelecida para determinar a correlação que tanto valida a *He1* como a *He2*;
- $r_i \Rightarrow$ ***posto*** *(1: mais similar ... 10: menos similar) do par variante/referente;* variável independente e ordinal, estabelecida para determinar os índices não-paramétricos do modelo estocástico, no cálculo de W e R_s)
- $W \Rightarrow$ ***correlação de concordância de Kendall***; variável dependente e de razão do modelo estocástico ($\in \Re$), estabelecida para determinar a correlação entre as *n* respostas dos indivíduos às perguntas do questionário e que varia no intervalo real [0, 1]. A Equação 6.4 formaliza o seu cálculo;
- $R_s \Rightarrow$ ***correlação de Spearman*** *entre os valores de SS (z) obtida pelo PRONTO© para cada par variante/referente, ordenados por postos, e os postos (r_i) obtidos das respostas ao questionário;* variável dependente e intervalar do modelo estocástico ($\in \Re$), estabelecida para determinar a correlação entre os resultados do protótipo e as respostas dos indivíduos às perguntas do questionário e que varia no intervalo real [-1, 1]. A Equação 6.5 formaliza o seu cálculo.

O critério para a avaliação do modelo de SS implementado por este protótipo usará como referencial a capacidade humana de julgamento da similaridade semântica de entidades[227] espaciais (geográficas) no mundo-real. Uma comparação de base estatística será delineada na metodologia.

[227] O mundo-real está "povoado" de *entidades*. As bases de dados ou de conhecimentos que representam esta realidade empírica estão "povoadas" de *objetos*.

6. A METODOLOGIA DE PESQUISA (plano do experimento)

"Em ciência, o que diferencia o pesquisador genuíno do fanático dogmático é que o primeiro é um humilde peregrino pela busca da verdade, enquanto o segundo considera-se o proprietário da verdade, porque foi iludido pela simples posse de uma certeza." (HUISMAN, 1976, p.11)

6.1. Considerações gerais

Inicialmente, a intenção que norteava o projeto desta pesquisa era a de enquadrá-la como um estudo notadamente experimental, dando continuidade explícita às concitações a trabalhos futuros propostas por RODRÍGUEZ[226] (2000), especialmente a que exorta futuros pesquisadores a comparar o seu MSS, de base ontológica, com modelos capazes de "desvendar" a semântica implícita nos modelos conceituais de bases de dados. A autora não deixou bem claro nessa proposta a qual tipo de base de dados se referia. Em correspondência a ela dirigida, (RODRÍGUEZ, 2001a) respondeu explicando tratar-se tanto de modelos mais tradicionais, como o relacional, assim como os mais "semânticos", orientados a objeto.

Para explicar o porquê da alteração do plano inicial de uma pesquisa que seria em bases experimentais para uma de natureza exploratória, é preciso apresentar um relato circunstanciado de eventos de ordem logística, que muito contribuíram para isso.

A pedido, RODRÍGUEZ (2001b) enviou o código-fonte (em C++) de seu MSS. No entanto, na mesma correspondência, explicava que o código, apesar de escrito numa linguagem de alta compatibilidade como o C++™[227], fora totalmente preparado para a plataforma *Macintosh*®[228], cujo sistema operacional é distinto do *Windows*® em termos de interface gráfica, necessitando de certas adaptações.

Como a experiência deste pesquisador na área da Ciência da Computação tende mais para o nível de análise do que para o de programação, foi tentada a via de se localizar algum programador com experiência em C++™ e na plataforma *Macintosh*®, confirmando-se o que já se pressentia: a pouca cultura nessa linha de sistemas operacionais no Brasil.

Os programadores procurados aconselharam a adoção de uma linha alternativa (mas onerosa), que seria a compra ou mesmo o aluguel de uma máquina da linha *Macintosh*® pa-

---

[226] p. 137 a 141 do original.
[227] C++™ consistente com o padrão internacional ISO/IEC 14882:1998.
[228] Macintosh® da Apple Computer, Inc.

ra a pesquisa. Essa linha de ação mostrou-se inviável pelo aspecto orçamentário, não dotado para esta pesquisa, havendo o agravante de se ter que contar com o equipamento por um longo período e não apenas para momentos pontuais do cronograma.

Com o aprofundamento da revisão de literatura, foram descobertas soluções mais simples[229] para a avaliação de SS, algumas bem minuciosas em termos de descrição algorítmica. Como algumas dessas descrições coincidiam com muitas estruturas de dados já implementadas numa emergente LTP – *Java*™ -, partiu-se para o estudo desta LTP, já que o pesquisador possui alguns conhecimentos na linguagem "C"™[230], da qual originaram-se o C++™ e o próprio *Java*™, ao mesmo tempo que se tentava resolver o problema da codificação de um protótipo que integrasse todos os conhecimentos acumulados na revisão de literatura, para conceber um modelo de avaliação de similaridade semântica de entidades espaciais.

Oportunamente, foi possível inaugurar uma linha de pesquisa num estabelecimento de ensino em que o pesquisador lecionava Tópicos Avançados em Programação (TAP)[231], aglutinando alunos que programavam profissionalmente em *Java*™. Ao aliar os objetivos de ensino de TAP, na fase de trabalhos em grupo, foi possível concitar os alunos a participarem parcialmente das fases da pesquisa que poderiam ser implementadas, atestando o seu modelo para o cálculo da SS numa *coleção de dados orientada a programas* (CDOP), montada com base nas definições que foram depois criadas para esta avaliação e para o questionário que foi aplicado a um grupo de técnicos de *geoprocessamento*.

Grifou-se acima CDOP para informar a intenção de tornar operacional o desenvolvimento do PROFAX. Até então, falou-se de base de dados da carta da região de FAXINAL como fundamento de aplicação para o PROFAX. Cabe, nesta altura, fazer a necessária distinção entre uma base de dados construída para servir de repositório para aplicações de um SGBD de uso geral ou específico. No caso em pauta, o SGBD que está por trás do projeto Faxinal é novidade no mercado – trata-se do SGBD do sistema *Gothic*®.

Rigorosamente, o *Gothic*® é um conjunto de aplicativos que utiliza dados armazenados num banco de dados (espaciais) orientado a objetos. O *Gothic*® está incluído num sistema mais abrangente de mapeamento orientado a objeto, denominado *Gothic*-LAMPS2™, da empresa de origem britânica LASER-SCAN, que faz parte do conglomerado de empresas anglo-americanas YEOMAN.

---

[229] Relatadas por RODRÍGUEZ (2000).

[230] C™ consistente com o padrão internacional ANSI/ISO-IEC 9899-1990 [1992].

[231] A ênfase da disciplina está em estudar os rudimentos da IA, no Curso de Sistemas e Informação da Faculdade Alvorada (Brasília, DF)..

O sistema básico LAMPS2™ foi originalmente projetado para atender às necessidades de mapeamento do Serviço de Hidrografia da Marinha Real Britânica, baseado num sistema de cartas eletrônicas navais[232], cujos dados (padrão S57[233]) seguem as especificações da IMO (*International Maritime Organization*) e da IHO (*International Hydrographic Organization*).

Atualmente, o conglomerado mencionado subscreve o OGC e participa de todos os esforços de interoperabilidade entre SIGs. Um desses esforços materializou-se na concepção de uma linguagem de programação para mapeamento em ambiente distribuído (*web mapping*): a GML™ (*Geographic Markup Language*).

Tais esforços em interoperabilidade têm colaborado na expansão da plataforma LAMPS2™ em diversos serviços de mapeamento do mundo: no Brasil, a DSG é a mais recente usuária; na Nova Zelândia, o serviço governamental de mapeamento desse país também já aderiu a essa TI; o serviço federal de geologia norte-americano (USGS) também utiliza essa tecnologia para estudos e mapeamento hidrográfico, só para citar algumas das organizações governamentais de mapeamento que aderiram à TI LAMPS2™.

A modelagem de dados orientada a objetos possibilita a obtenção de uma melhor aproximação do mundo real, acrescendo-se "inteligência" aos dados e aumentando a confiabilidade dos produtos finais do sistema (cartas, imagens, mapas, etc.).

A capacidade de atribuir comportamentos a um objeto é fundamental para se definir como ele influenciará o meio que o envolve, caso sofra alterações (LUNARDI, 2001). O modelo orientado a objetos é, portanto, dinâmico, em oposição ao relacional.

Essa capacidade oferecida pela OO fornece a um SIG uma vantagem inatingível em relação a um sistema baseado num modelo relacional de banco de dados, incapaz de herdar atributos de uma classe superior por herança. Essa vantagem dos sistemas OO agiliza algumas das fases do mapeamento sistemático em gabinete, como a edição e também reduz sensivelmente o espaço de armazenamento (memória) para dados cartográficos digitais.

Durante a exposição de LUNARDI (2001), acerca do projeto-piloto[234] que a DSG vem implementando com base na carta Faxinal, já se pôde identificar algumas facetas metodológicas de produção inusitadas em relação ao mapeamento em vigor, ainda baseado na CAC. Em particular, o SGBD espacial cativo do *Gothic*® permite uma auto-estruturação dos dados já carregados nas bases de dados espaciais, atualizando com rigor lógico-geométrico o item mais crítico para SIGs apoiados nos SGBDs relacionais: a topologia[235]. Trata-se do recurso

---

[232] ECDIS: *Electronic Chart Display and Information System.*
[233] Special Publication 57, edition 3.0 (LASER-SCAN, 2002).
[234] Inserido no plano maior (PCE), de representar todo o espaço geográfico brasileiro num SIG de grande envergadura.
[235] V. glossário.

on the fly©, em que não é necessário refazer o arquivo de topologia sempre que os dados espaciais são manipulados. É um enfoque dinâmico para atualizar bases de dados e não estático e parcialmente dependente do operador humano, como no modelo relacional.

Como se viu da complexidade tecnológica e do vulto (governamental) dos investimentos numa TI como a do sistema *Gothic*-LAMPS2™, nesta pesquisa não se poderia nutrir a pretensão de trabalhar numa plataforma tão complexa para fornecer o mínimo de evidência empírica ao que constitui o seu objetivo geral: ***"Avaliar a similaridade semântica entre objetos espaciais representáveis numa base de dados***[236]***"***.

Nem para uma equipe de produção e desenvolvimento de *software* seria possível preparar em apenas um ano[237] um protótipo nos moldes do PRONTO©, que fosse capaz de se vincular à verdadeira base de dados (carta Faxinal) do sistema *Gothic*-LAMPS2™.

Em primeiro lugar, a equipe lá alocada pela 1ª DL (1 analista sênior e quatro programadores[238]) para iniciar a execução do **PMEGB** (Projeto de Modelagem do Espaço Geográfico Brasileiro) levou cerca de dois anos apenas para operar a capacidade mínima de produção do sistema *Gothic*®.

Em segundo lugar, esta pesquisa não tem seu esforço concentrado no aspecto tecnológico, mas no científico e, portanto, qualquer simplificação plausível do primeiro aspecto pelo princípio da ***Navalha de Ockham*** (consulte o glossário), para sustentar um pressuposto ou hipótese, deveria ser explorada.

Em terceiro lugar, esta pesquisa acabou por se ajustar à tipologia: exploratória e deverá desenvolver idéias, refinar conceitos e enunciar questões e hipóteses para investigações subseqüentes.

Segundo TRIPODI (1975), numa pesquisa desse gênero, uma variedade de procedimentos de coleta de dados pode ser usada, porém menos atenção deve ser devotada à descrição exata de relações quantitativas entre variáveis, distintamente dos estudos quantitativo-descritivos. E foi justamente nesse aspecto que a pesquisa entrou em mutação, porque o seu caráter abrangente e o estado-da-arte que estava se montando não justificavam mais o foco num experimento que seguisse o rigor dos métodos experimentais.

O fenômeno da SS é também de natureza subjetiva e o que mais se deseja nessa linha de pesquisa, iniciada na Universidade do Maine, é montar um repertório vasto de conhecimento sobre o assunto, de modo que se possa implementar, aqui e ali, uma comodidade de

---

[236] Lembrar que nos objetivos específicos foi apresentada a verdadeira dimensão de base de dados para os protótipos.
[237] Tempo concedido pelo órgão de origem do pesquisador, exclusivamente dedicado à elaboração da tese.
[238] Também insuficiente em quantidade de profissionais para a envergadura do projeto.

comunicação homem-máquina. Essa comodidade não pode ser iniciada por pesquisas de pesadas características quantitativas, até porque não há elementos concretos que possam ser transformados em números. Será que uma instância de classe é melhor individualizada pelos parâmetros de partes, funções e atributos determinados pela Psicologia Cognitiva? Será que a recente entrada dos cientistas da computação nesses problemas não desvendará outras formas de avaliar a SS, de maneira mais objetiva? Há de se convir que só um estudo de natureza exploratória pode dar alternativas (e não respostas definitivas) para essas questões.

TRIPODI (1975) argumentou que esses estudos incluem uma grande quantidade de informações para um estudo-de-caso, mas são, contudo, menos definidos que os estudos experimentais e quantitativo-descritivos, nos quais se tenta associar variáveis ou verificar hipóteses.

Estudos exploratórios de manipulação experimental são aqueles estudos exploratórios que manipulam uma variável independente, a fim de localizar variáveis dependentes que estejam potencialmente associadas à variável independente. Caracteristicamente, uma unidade de comportamento é estudada em seu meio natural. Freqüentemente, o propósito desses estudos é demonstrar a viabilidade de um determinado programa ou técnica como uma solução em potencial para problemas práticos. Uma variedade de procedimentos de coleta de dados pode ser empregada e técnicas de observação podem ser desenvolvidas durante o transcurso da pesquisa (TRIPODI, 1975).

Como se depreende da leitura do parágrafo anterior, que fornece a definição completa de um estudo exploratório de manipulação experimental, depois de todas as características até aqui apresentadas sobre esse estudo de caso de SS, não resta dúvida que o enquadramento nesse tipo é muito adequado. Até pelo público-alvo da obra de TRIPODI (1975), que pertence à área da Psicologia experimental, a definição e os exemplos complementares que constam do texto vão ao encontro do tema desta pesquisa.

Uma outra particularidade de estudos exploratórios, que se coaduna com esta pesquisa, é que o processo da descoberta não é suficientemente enunciado para que o pesquisador possa seguir um conjunto prescrito de regras. Na verdade, tal processo criativo não segue necessariamente regras metódicas de lógica. É o que se configura com o material de trabalho desse estudo-de-caso, com alguma carga subjetiva, segundo a maioria dos autores da área das ciências cognitivas compulsados no Capítulo 3. Entretanto, neste trabalho, são acatadas as recomendações de estudiosos que estruturaram o processo investigativo, de tal modo a torná-lo mais robusto, i.e., para que a probabilidade da descoberta seja aumentada.

Essas diretrizes metodológicas aplicam-se em três etapas do processo exploratório de pesquisa: Fontes de informações, tipos de dados e uso de dados.

Agora, uma ênfase sobre a primeira etapa acima: fontes de informações.

Como se viu na revisão de literatura, são várias as citações de autores e pesquisadores, de autoridade reconhecida no âmbito científico ao qual pertencem, abonados nas referências bibliográficas por meio de registros não tão corriqueiros[239], como respostas por correspondência eletrônica e até conversas pessoais. Daí exsurge um problema de vulto das pesquisas exploratórias, que não deixa de estar presente neste trabalho: a sobrecarga de dados e informações.

Para assimilar grandes quantidades de informações de natureza qualitativa, em porções manipuláveis ao seu entendimento, pode ser necessário aplicar técnicas de *análise de conteúdo*, *análise fatorial* (*factor analysis*) ou *análise de componentes principais (principal component analysis* – PCA). Pode ser interessante, num estudo futuro, um aprofundamento na teoria sobre análise fatorial e PCA, em que são referências RUMMEL (2002), HYVARINEN[240] (2002) e UCN[241] (2002).

Com relação à amplitude que o assunto SS pode ser tratado, dando mais convicção de que a matéria pertence ao domínio de interesse das pesquisas exploratórias, citam-se alguns trabalhos que aplicaram a formulação do co-seno entre dois vetores no espaço euclidiano e técnicas de PCA para resolverem problemas de seus respectivos campos de conhecimento:

- Estudo de distribuição de espécies de delfinídeos na costa atlântica da América do Sul (BARRETO, 2000);
- Experimentos de esporulação de fermentos (RAYCHAUDHURI, 2002);
- Medidas de emissões de motores a *diesel* (McADAMS, 2002).

Este trabalho está assentado na mesma linha do de RODRÍGUEZ (2000), que aplicou seu esforço principal de pesquisa na elaboração de um sistema conceitual que fosse tratado por um sistema computacional. Esse esforço traduziu-se num movimento de desenvolvimento de cima para baixo (*top-down*), em termos de Engenharia de *Software*, não se preocupando com pormenores em níveis mais baixos de projeto, implementação de aplicativos ou mesmo de estruturação de dados armazenados em bases de dados.

---

[239] Obras consolidadas em livros, revistas, etc.
[240] Centro de Pesquisa em RNA da Universidade Técnica de Helsinque.
[241] Univ. da Carolina do Norte: lista de discussão eletrônica PA765 (moderador: David_Garson@ncsu.edu).

Foi justamente para explorar a semântica "escondida" nas estruturas de dados que povoam as bases de dados, que RODRÍGUEZ (2000) teve a idéia de concitar um trabalho nessa área e que a presente pesquisa procura explorar parcialmente.

Em síntese, este estudo exploratório foi baseado na pressuposição de que por meio do uso de uma metodologia híbrida, entre métodos quantitativos (MM do MSS e estimadores estatísticos) e qualitativos (análise e interpretação dos resultados numéricos, etc.), é possível desenvolver hipóteses relevantes para a SS.

6.2. Campos de observação

"A causa está oculta, mas o resultado é conhecido."
[Ovídio (*apud* WONNACOTT (1980, p. 283)]

Em termos de montagem do *corpus*, esta pesquisa seguiu uma linha metodológica aproximada entre MEDEIROS (1999) e RODRÍGUEZ (2000).

O acervo de dados utilizado por MEDEIROS (1999) foi constituído de textos em língua portuguesa completos, armazenados nas bases de dados do Mercosul, que são mantidas pelo Ministério das Relações Exteriores. É sobre este acervo que foram realizados testes de tratamento de ambigüidade lingüística. A aplicação de um sistema especialista (Zstation™), baseado nas regras de PLN da IA, foi o instrumento básico para a autora construir a sua base de conhecimentos, a fim de detectar a ambigüidade inerente à fonte lexical.

No caso de RODRÍGUEZ (2000), seu *corpus* foi montado com base em duas linguagens documentárias: SDTS™ e *Wordnet*™.

SDTS é uma sigla que significa *Spatial Data Transfer Standard*, um catálogo especializado de termos espaciais, criado pelo serviço geológico norte-americano (USGS). As entidades espaciais desse catálogo são expressas por termos grupados em classes e padronizados segundo as normas do Instituto Americano de Padrões (ANSI) para o Intercâmbio de Dados. Este produto indica que os esforços norte-americanos estão numa fase bem mais adiantada do que os esforços brasileiros do CEPAD/CONCAR, hoje em dia, descontinuados.

O SDTS contém cerca de 200 classes e 1200 subclasses de termos, suas relações de gênero-espécie (é-um) e seus atributos.

A abreviatura *Wordnet* refere-se a uma classificação disponível na Internet, desenvolvida no laboratório de ciência cognitiva da Universidade de Princeton. Esta classificação *on-line* é um sistema de conceitos organizado em conjuntos de termos sinônimos (*synsets*), in-

ter-relacionados semanticamente. São cerca de 118.000 termos, organizados em 90.000 conjuntos de sinônimos.

A *Wordnet*™ é uma linguagem documentária muito usada em sistemas de controle de ambigüidade de textos (*desambiguação*), na recuperação de informação e modelagem conceitual de sistemas de informação que operem com relações semânticas [(RODRÍGUEZ, 2000) e (CLUL, 2002)].

O trabalho de RODRÍGUEZ (2000) combinou o catálogo SDTS™ com a classificação *Wordnet*™ na construção do *corpus* para carregar no protótipo de implementação do seu modelo conceitual – o MSS. Os *synsets* extraídos da *Wordnet*™, acrescidos de relações de gênero-espécie ou hiponímicas (é-um) e de relações de composição ou meronímicas (todo-parte) complementaram as definições dos tipos de classes do SDTS™.

Para avaliar o seu *corpus* e o protótipo sobre ele aplicado, RODRÍGUEZ (2000) se utilizou de um referencial humano para a fase de testes. Portanto, a autora trabalhou em dois[242] campos de observação: o do ambiente computacional e o do ambiente cognitivo dos indivíduos que participaram da pesquisa.

Tratando-se desta pesquisa, o caminho seguido não foi muito diferente do adotado pela autora até aqui considerada. O acervo de dados cartográficos[243](*corpus*) foi retirado primordialmente de três fontes:

- Tabela da base de dados cartográficos digitais (TBCD©) da DSG;
- Manual técnico T34-700© da DSG;
- Modelo conceitual do espaço geográfico brasileiro, materializado pelo projeto-piloto da carta Faxinal (PR).

Quando as fontes primordiais se mostraram insuficientes para construir as definições dos termos que foram utilizados para o cálculo da SS e para colher respostas dos indivíduos (referencial) do CCAuEx, foram utilizadas fontes alternativas como: vocabulários controlados (*Wordnet*™ ou similar), dicionários enciclopédicos (LELLO, 1984), livros de geociências (LEINZ, 1980), glossários (BRASIL, 1985) e obras de terminologia [(GARCIA, 1976), DAHLBERG, 1978), (FELBER, 1984)].

Neste ponto é bom fazer uma nota, visto que, durante a revisão de literatura, a única taxinomia *on-line* em língua portuguesa ainda estava nos seus primórdios, a versão 1.0 da *Wordnet*.PT™, de responsabilidade do Centro de Lingüística da Universidade de Lisboa

---

[242] Os mesmos estabelecidos para este trabalho (V. Figura 6.1).
[243] Temática variada: hidrografia, relevo, infra-estrutura, localidades, vegetação, etc.

(CLUL, 2002), o que aumentou a carga de trabalho na consulta a fontes alternativas para montar as definições dos termos com o maior rigor lógico e sistematização possíveis.

A Figura 6.1 dá uma idéia dos dois campos de observação desta pesquisa.

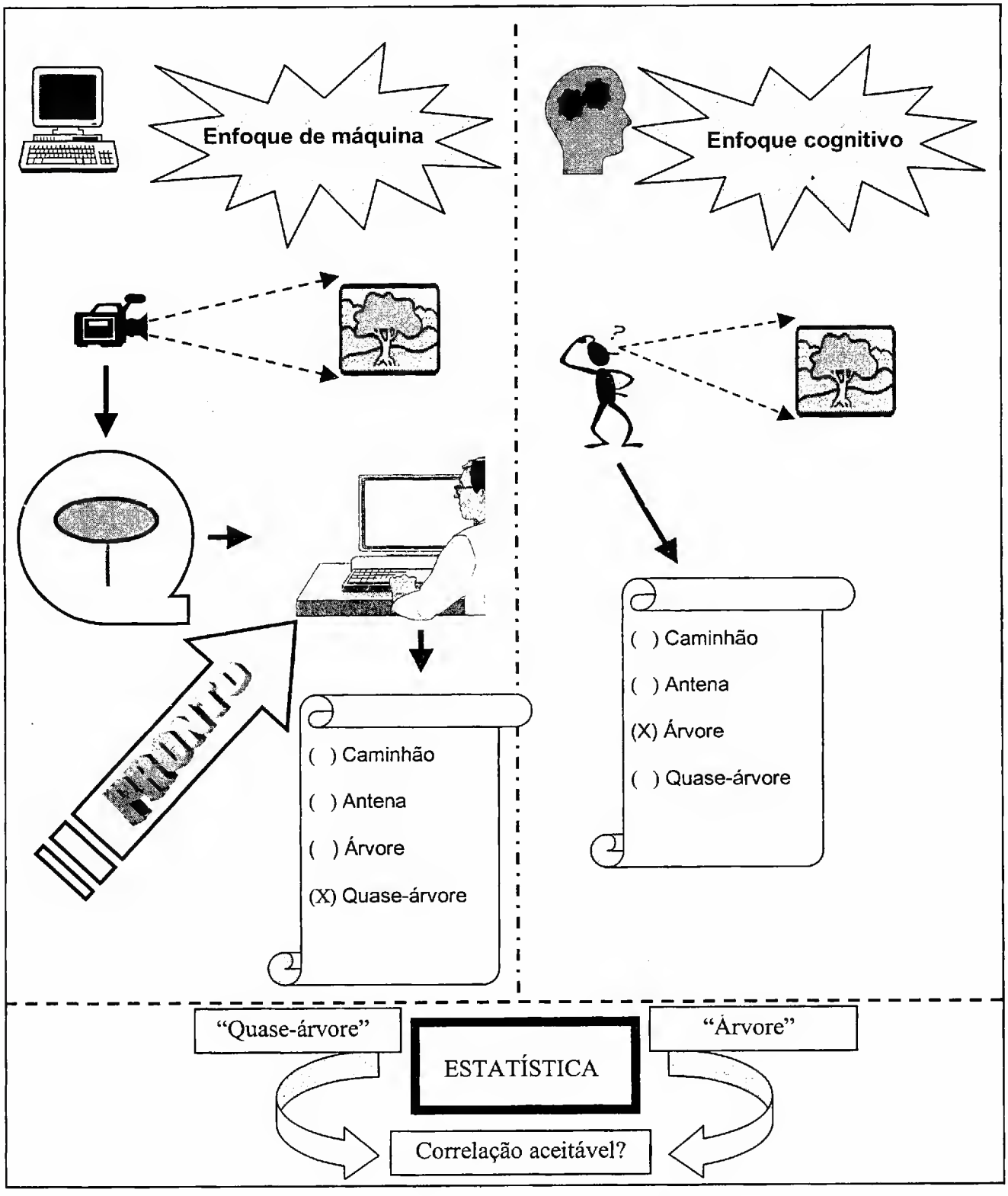

Figura 6.1: Campos de observação da pesquisa.

6.3. Instrumentos (meios) de pesquisa

"Os modelos existem para ser usados, não para acreditarmos neles."
[Henri Theil (*apud* WONNACOTT (1980, p. 297)]

Os dois instrumentos de pesquisa utilizados nesta tese para validar o modelo se avaliação de SS são um protótipo de *software* e um questionário. O primeiro será avaliado tendo por referencial o resultado do segundo.

O protótipo foi desenvolvido em duas fases. A primeira foi destinada para tomar contacto com um modelo matemático muito disseminado e utilizado de SS – a fórmula do co-seno – e para produzir o repertório de conhecimento necessário para a implementação de uma ferramenta mais aprimorada desse programa, o PRONTO©.

A técnica básica de utilização do modelo matemático da SS pelo co-seno e congêneres (*Dice Coefficient*, Coeficiente de Jaccard[244], Distância Euclidiana e Distância de Manhattan), tratados por WONG (2000), GANESAN (2002), SANTOS (2002), FETI (2002) e DCCUT (2002), é a dos exemplos de treinamento (*training examples*), empregados na automação da classificação documentária e da recuperação de informação textual. Da mesma forma, a técnica de *exemplos de treinamento* foi a utilizada para demonstrar[245] o êxito do PROFAX.

No caso da segunda ferramenta - o questionário -, em linhas gerais, seu preparo e aplicação seguiram em boa parte a estrutura adotada por RODRÍGUEZ (2000) e outras recomendações complementares, como a do cientista David Mark (MARK, 2002), que aconselhou a não usar qualquer tipo de subsídio material, como a carta de Faxinal (PR) impressa ou um modelo tridimensional do terreno desta região, como se desejava, para apoiar os indivíduos respondentes na solução do questionário, visto que esta ação produziria ruídos de natureza desconhecida nos resultados, ligados ao efeito de variáveis intervenientes no experimento, difíceis de controlar.

As diferenças no quantitativo amostral foram muito pequenas. No caso de RODRÍGUEZ (2000), o seu questionário (5 perguntas) foi aplicado a 72 estudantes do Curso de Letras da Universidade do Maine, em que estes deveriam pôr em ordem de SS (de 1 a 11, sendo 1 o mais similar) pares de termos que se referiam a objetos espaciais, retirados pela pesquisadora de sua ontologia, montada essencialmente com base em dois vocabulários-taxinomias controlados (*Wordnet*™ e SDTS™), como já explanado.

---

[244] Alguns autores (DCCUT, 2002) grafam Jacquard; outros (GANESAN, 2002), Jaccard.
[245] A demonstração bem-sucedida se deu no exame de qualificação, em 4 de setembro de 2002.

No caso da presente pesquisa, foi aplicado um questionário similar (6 perguntas) para 67 técnicos da DSG e do CCAuEx, em que estes deveriam pôr em ordem de SS (de 1 a 10, sendo 1 o mais similar) pares de termos que se referiam a objetos espaciais retirados de uma ontologia *ad-hoc*, construída essencialmente com base no MC (modelo conceitual) da folha Faxinal.

No entanto, as diferenças de ordem qualitativa são grandes, já que o questionário de RODRÍGUEZ (2000) foi aplicado a estudantes universitários de um curso de ciências humanas[246] e homogeneamente divididos segundo o sexo, enquanto que, no caso em tela, não houve ocorrência de indivíduos do sexo feminino, todos os respondentes são habilitados em técnicas muito específicas de *geoprocessamento*, a maior parte dos indivíduos não possui curso superior, entre outras particularidades que virão à tona na seção de análise dos resultados.

Apurados os resultados pelos dois instrumentos de pesquisa, sucedeu-se a fase dos testes estatísticos apropriados, consoante a mesma metodologia usada por RODRÍGUEZ (2000)[247], que justifica a aplicação de testes não-paramétricos: 1) O coeficiente *W* de *Kendall,* para as correlações entre as respostas dos indivíduos e 2) O coeficiente de *Spearman,* para correlações entre as respostas dos indivíduos e os resultados do PRONTO© [(DANIEL, 1978, p.326), (SIEGEL, 1988, p. 262) e (MILLER, 2002, p.217)].

6.3.1. Protótipo - generalidades

Como anunciado no subitem 1.3, este pequeno espaço destina-se à descrição superficial do que seja um protótipo, pelo enfoque de um dos ramos de aplicação da Ciência da Computação, a Engenharia de *Software*, preocupada com a aplicação dos métodos de engenharia no desenvolvimento e na manutenção de *software* (PRESSMAN, 1995).

Para se pensar em construir um SI (sistema de informação), é preciso, antes, construir um modelo desse sistema. No entanto, há situações em que o cliente (sistema-usuário), mesmo delineando os objetivos gerais que deseja para o seu sistema, ainda possa não ter identificado os requisitos de entrada, processamento e saída; nem tampouco o desenvolvedor ainda tenha certeza da eficiência de um algoritmo, da adaptabilidade de um sistema operacional ou da forma em que se deva dar a interação homem-máquina. Esses são os ingre-

[246] Sem afinidade com a área *geocientífica* (curso de Letras).
[247] p. 92 a 98 de RODRÍGUEZ (2000).

dientes indicados para se adotar um dos quatro paradigmas[248] (ou enfoques) da Engenharia de *Software*; e o que mais se apropria para esta situação de incertezas é a **prototipação**.

A prototipação inicia-se com a coleta de certos requisitos e, daí, se passa para uma fase chamada de *projeto rápido*, ou melhor, uma fase em que o usuário (organização) e o desenvolvedor (equipe ou organização) concentram-se nos aspectos do *software* que serão visíveis ao primeiro (enfoques de entrada e formatos de saída). Em seguida, resulta o protótipo, i.e., a chave para uma "sintonia fina" entre o usuário e o desenvolvedor, porque o primeiro consegue uma forma gradual de ir atendendo às suas necessidades e o segundo tem a oportunidade de ir compreendendo melhor aquilo que precisa ser feito no futuro sistema.

Segundo PRESSMAN (1995), a Engenharia de *Software* já atingiu um certo grau de maturidade, ao pôr à prova, por cerca de quatro décadas, os seus métodos e técnicas, para que a ordem *protótipo – modelo – sistema* não seja subvertida, caso as incertezas já citadas estejam presentes.

Esta pesquisa foi norteada por essa cautela, especialmente por estar tratando de maneira direta (quando é mais engenharia) ou indireta (quando é mais Ciência da Informação) desse produto intangível do intelecto humano, chamado ***software***.

A *prototipação* é uma metodologia voltada à aceleração do desenvolvimento de sistemas, com a participação ativa do usuário, para implementar um **protótipo de trabalho** que execute algum módulo funcional de um sistema futuramente desejado ou que realize parte da função de algum sistema já existente e que tenha outras características que não possua esse sistema, mas que serão necessárias e que poderão ser melhoradas num esforço de desenvolvimento futuro (PRESSMAN, 1995).

Como se vê, um protótipo serve como um mecanismo que auxilia a identificação dos reais requisitos de *software* e, assim, ajusta-se à justificativa que foi dada para conduzir este estudo exploratório, i.e., contribuir para o desenvolvimento de futuros SIGs interoperáveis.

O PROFAX e o PRONTO© são protótipos de trabalho, i.e., programas que implementam algumas funções da relação de SS para:

- Percorrer a estrutura de dados hierárquica (árvore *n-ária*) que contém as classes de objetos espaciais para estudo;
- Capturar (recuperar) os nós relevantes para um determinado cálculo de SS;
- Calcular a SS entre os nós selecionados.

---

[248] Os outros três são: o ciclo de vida clássico, o modelo espiral e as técnicas de 4ª geração.

Reitera-se a restrição já declarada de que não se pretende construir[249] um sistema ou um módulo (subsistema) de sistema de informação nesta tese. Sendo assim, de forma geral, o PROFAX e o PRONTO© não deixam de incorporar as seguintes características típicas de um protótipo (PRESSMAN, 1995):

- Rapidez na produção de sistemas de informação;
- Suporte à modelagem de sistemas de informação;
- Flexibilidade;
- Envolvimento do usuário;
- Utilização de linguagens de 4ª geração (*Java*™, no caso);
- Mais adequado aos sistemas não-estruturados (difíceis de formalizar), que são mais comuns em níveis bem altos (estratégicos) de uma organização.

Entre os objetivos já delineados para esses protótipos, no subitem 1.3, a construção desses *softwares* ainda engloba os seguintes objetivos gerais (PRESSMAN, 1995):

- Reduzir os prazos de desenvolvimento dos sistemas de informação;
- Aumentar a produção de informações desses sistemas;
- Garantir a participação do usuário;
- Reduzir os custos de manutenção dos sistemas de informação.

Pelas características e objetivos gerais enumerados para um protótipo, é fácil apurar que se trata de uma ferramenta perfeitamente adequada para o desenvolvimento de um estudo exploratório, que se insere na tipologia de ***manipulação experimental***, segundo TRIPODI (1975).

Finalizando a descrição sobre protótipo, pelo enfoque das aplicações em Ciência da Computação, não é demais enumerar os riscos que o abuso na prática da *prototipação* pode introduzir no desenvolvimento de sistemas. Por conseguinte, a literatura sobre Engenharia de Sistemas e de *Software* alerta para os seguintes riscos, na ordem decrescente de importância para pesquisas de cunho mais científico do que tecnológico:

- Descontrole na documentação dos sistemas;
- Dificuldade para a integração de subsistemas;
- Acomodação do usuário com soluções expeditas para os seus problemas;
- Proliferação de produtos na área de computação de uma organização;
- Crescimento descontrolado dos custos de manutenção de *software*.

---

[249] Mas auxiliar esta construção no futuro.

Com respeito aos dois primeiros riscos da enumeração anterior, cabe salientar que por ser este um trabalho de pesquisa realizado num curso de CI, em que a documentação do conhecimento é um ponto de honra, foi dispensada especial atenção para esse item, cujo resultado somente a continuação do trabalho poderá comprovar.

No caso de integração de subsistemas e da acomodação do usuário a soluções de fortuna e eventuais, já foi exaustivamente comentado o fato da inserção deste trabalho de pesquisa num quadro referencial de amplas perspectivas no campo da Ciência da Informação Espacial, não merecendo maiores justificativas sobre o pouco risco que esse item representa para o trabalho.

Com relação aos dois riscos remanescentes, somente a visão mercadológica que esses desenvolvimentos podem suscitar é que ditará a consistência ou inconsistência dos protótipos que se produzem nesse ramo do conhecimento, não sendo do interesse imediato desta pesquisa e das que lhe são congêneres atender ao teor desse enfoque.

A Figura 6.2 ilustra as fronteiras que a *prototipação* impõe entre o usuário (organização), o sistema em foco (produto) e o desenvolvedor ou construtor de *software* (processo).

Resumindo a noção de protótipo numa frase de PRESSMAN (1995, p. 1.011):

"Uma idéia é formulada e evolui para um protótipo, que é usado para demonstrar conceitos básicos."

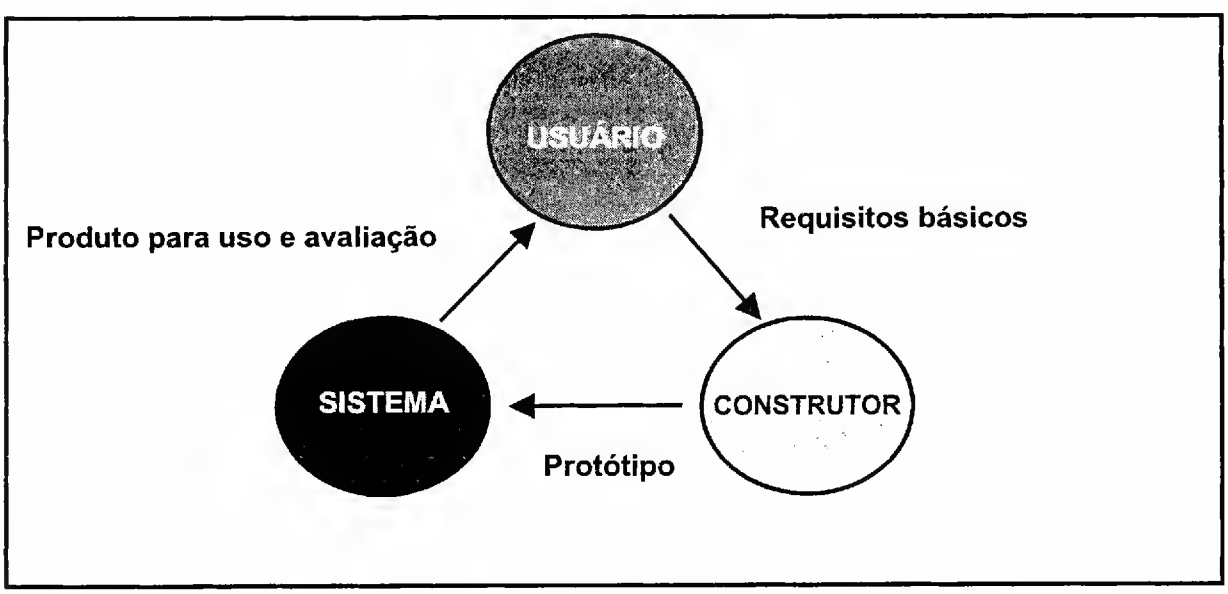

Figura 6.2: Os elementos da prototipação.

6.3.1.1. Primeiro protótipo para um modelo vetorial de avaliação de similaridade semântica (PROFAX)

Esse protótipo denominou-se PROFAX (PROtótipo de avaliação da similaridade semântica para as classes de entidades espaciais da carta de FAXinal), muito importante para avaliar modelos de cálculo da SS que utilizam conceitos da Álgebra Linear e para recuperar objetos armazenados em bases de dados.

A idéia básica foi construir pares de vetores no espaço *n-dimensional* euclidiano, em que cada vetor representou um pacote de termos referentes a classes de entidades espaciais. Um pacote foi o vetor que continha os termos de busca; o outro, o vetor que continha os termos mais significativos do repositório de dados sobre o qual a busca foi feita.

A seguir, sobre ambos os vetores do par, foi aplicada a fórmula do produto interno (co-seno entre dois vetores), para obter resultados que, devidamente apropriados à Teoria da Similaridade (TVERSKY, 1977) e suas restrições, foram definidos como indicadores da similaridade entre os vetores de cada par. A variação desses indicadores ficou entre **zero** (busca malograda, i.e., similaridade mínima entre os vetores do par) e **um** (busca bem-sucedida, i.e., similaridade máxima entre os vetores do par).

A seguir, será exposto um memorial resumido da metodologia de construção do PROFAX, por linhas de ação (duas delas).

A **1ª linha de ação** foi o desenvolvimento de um programa (em *Java*™) para armazenar os termos e feições dos objetos espaciais representados no MC da carta topográfica de Faxinal (PR), elaborado pela **1ª DL**. Alguns tópicos relevantes dessa linha:

- *Corpus*: por inspeção visual, sem qualquer critério, foram extraídos alguns acidentes naturais e artificiais do MC/Faxinal (notação gráfica da UML™).
- Função de cálculo da similaridade semântica (SS): foi escolhido o produto interno entre dois vetores (co-seno), com base nos trabalhos de GANESAN (2002), SANTOS (2002), FETI (2002), DCCUT (2002) e WONG (2002). O aporte conceitual para montar as estruturas de relações (de generalização e de agregação) foi tirado de FELBER (1984, p. 178 e 227), GARCIA (1976) e da tese de doutorado de RODRÍGUEZ (2000). O modelo matemático desta última pesquisadora, híbrido entre o vetorial (distância semântica – que é a base dos cinco primeiros trabalhos citados neste parágrafo) e os estatísticos (mais complexos), seria difícil de reproduzir no tempo que era disponível para essa fase da pesquisa. Eis a razão de não se ter construído uma estrutura formal como uma ontologia para carregar nesse protótipo.

- Implementação: em parceria com um grupo de pesquisa da área de Sistemas de Informação. A estrutura de dados básica do programa que calcula a SS é a estrutura de árvore (*n-ária*), como ilustrado nas Figuras 6.3(a, b, c).

A **2ª linha de ação** foi a montagem (manual) de um exemplo simples de criação de definições, que complementou o *corpus* e serviu de referencial para comparar os resultados do PROFAX.

Na realidade, para montar esse exemplo, não foram definidos os termos que representam as entidades espaciais, mas apenas grupados esses termos, segundo suas feições distintivas (traços, características), numa árvore (taxinomia), para permitir o cálculo de SS pelo PROFAX. Essas feições distintivas (*fds*) foram obtidas por pesquisa em dicionários, manuais técnicos e na própria experiência do autor.

A extração dessas feições constitui uma das fases necessárias para produzir uma obra de Terminologia Científica [FELBER (1984, p.179) e RODRÍGUEZ (2000, p.107)]. Nessa fase, cada termo pode ser formulado segundo a expressão designada por *estrutura formal da definição predicativa* (GARCIA, 1976, p.308), vista na p. 146.

É importante lembrar que *propriedades* não possuem o poder de categorização que possuem as *características* [Gottlob Frege, *apud* PINTO (1977) e MODELL(b) (2001)]. Mas aqui, neste trabalho, para aclimar diferentes nomenclaturas, confundem-se ambos os termos com feições distintivas[250] (***fds***).

Frisa-se que para o teste do PROFAX não se extraíram as *fds* de entidades como rio, lago, estrada asfaltada, ferrovia, etc. Para simplificar e fazer o PROFAX funcionar nessa versão inicial, só se pôde trabalhar com coleções dessas entidades (p.ex: rio e lago, uma coleção; estrada asfaltada, ferrovia e trilha, outra coleção, etc.).

Entretanto, embutido no objetivo geral desta pesquisa, está a determinação de se separar as feições distintivas pelas definições, o que se fez na versão seguinte do protótipo.

Apesar de ser recente a exploração dos conceitos da *classificação científica* na Ciência da Computação, especialmente com a criação das linguagens de programação orientadas a objeto, as chamadas ciências da classificação (entre elas, a Biblioteconomia) há muito tempo, mesmo antes do advento de meios automatizados de processamento de dados, vêm acolhendo em seus objetos e metodologias um rico repertório de princípios que orientam a obtenção de conhecimento por esse *modus sciendi*, que é a classificação.

---

[250] RODRÍGUEZ (2000) as denomina de *distinguishing features*.

É bem conhecido na Ciência da Informação e na Biblioteconomia o potencial da análise por ***facetas***, i.e., várias hierarquias que podem ocorrer na classificação de uma área do conhecimento. A análise por facetas é um exercício de definição, em que é procedida uma análise conceitual de um assunto segundo as características essenciais das entidades que o compõem, de acordo com os requisitos do grupo de usuários que compartilham uma área do saber. Percebe-se uma clara semelhança dessa metodologia com os princípios que sustentam uma das teorias do conceito: a Teoria do Protótipo.

A definição das facetas pode ser orientada segundo aspectos essenciais do fenômeno: suas propriedades (P), comportamentos (C), interações (I) e operações (O): PCIO.

Segundo QUILLIAN (1968), esse mecanismo de classificação se dá pelo "disparo" de planos conceituais de maior conteúdo informativo, com base em planos mais genéricos. Essa é a base formal das redes semânticas.

Sendo assim, o quadrinômio PCIO e correlatos podem ser considerados como *constructos* fundamentais de categorização do conhecimento.

Foi Shialy Ramarita Ranganathan, teórico e bibliotecário hindu, que em 1967 lançou o conceito de *facetas*, embutido na idealização das suas "regiões convencionais do conhecimento", regularmente homogêneas, mutuamente exclusivas, fortemente hierarquizadas e totalmente completas. Ele procurou contribuir com a Teoria da Classificação de forma integradora (holística).

Segundo VICKERY (1975), as facetas de Ranganathan mostraram-se inadequadas, em boa parte devido ao aspecto demasiadamente abstrato das concepções do autor, mas nem por isso se deixou de estudar a sua obra fundamental até hoje. É bem provável que o grande impulso dado às linguagens documentárias (*tesauros*, p.ex.), em direção às tecnologias da informação, tenha tido sua origem na análise de um assunto por facetas.

Por se tratar de assunto notório e até de remoto conhecimento por parte dos estudiosos das ciências da classificação e pelas recentes explorações que sobre tal assunto se têm debruçado os pesquisadores da IA, é pertinente explorá-lo nessa região de fronteira de conhecimentos que se formou, em que a linguagem cartográfica vem produzindo novos conceitos e até mesmo materializando-os em bens e serviços, como é o caso das TBCD© (V. glossário e subitem 3.1.1).

Como já explanado (RUSSELL, 1995, p. 222), pelo enfoque da IA que foi adotado neste trabalho e que foi básico para a implementação do PRONTO©, a gênese de uma ontologia resume-se a uma lista de conceitos de um certo domínio do conhecimento. Ela deve conter

sentenças descritivas (*axiomas*) sobre os termos que designam esses conceitos. Escrevendo essas definições, cumprem-se dois objetivos da ontologia:

- O primeiro objetivo está ligado à delimitação (mais precisão) dos conceitos que povoam os espaços cognitivos dos indivíduos que comungam de um repertório lingüístico especializado (língua profissional), incrementando a comunicação e o entendimento entre esses indivíduos e amenizando o efeito indesejável da ambigüidade;
- O segundo objetivo está ligado às conseqüências de natureza formal que a lista citada acarreta. Dessa formalidade, um produto que inevitavelmente há de se formar é uma BC afetada por um mecanismo de inferência.

Como se verá na montagem das ontologias que definem as classes de entidades espaciais da folha Faxinal, que estão representadas numa base de objetos, o estabelecimento de hierarquias por facetas não foi menosprezado, já que levantar *fds* de uma classe de entidades espaciais nada mais é do que selecionar as características e comportamentos representativos (facetas) daquela classe, segundo categorias fundamentais do conhecimento.

A regra básica, com base em VICKERY (1975, 25-51), para a montagem de uma ontologia no âmbito da IA, i.e., com apoio conceitual nas técnicas de classificação, é começar a se definir as categorias mais genéricas. Dentro de cada categoria, então, passa-se a definir os termos designativos das classes subordinadas, dos atributos ou propriedades dessas classes e dos comportamentos verificados entre essas classes.

Para coisas e suas propriedades ou atributos, em primeiro lugar, deve-se preferir nomeá-las por substantivos, com prevalência para os comuns, concretos e simples (p.ex: *alcance*); em segundo lugar, nomeá-las por adjetivos substantivados (*alcançador*) e por último nomeá-las pelas formas verbais nominais (infinitivo, particípio e gerúndio: *o alcançar*). No caso dos comportamentos, a ordem é preferir verbos no infinitivo e nas formas concretas, com significação própria (*haver* no sentido de *existir*) e deixar para segundo plano ou nem usar as abstratas ou auxiliares (ALMEIDA, 1994, p. 241), que são vazias de sentido.

VICKERY (1975, p. 31-37) cita seis dessas categorias do conhecimento: **substância**, **estado**, **propriedade**, **operação**, **reação** e **instrumento** ou aparelho, um aprimoramento ao quadrinômio PCIO já citado.

Uma interessante tentativa de modelar uma área de assuntos cartográficos poderia ser imaginada pela construção de vetores com seis dimensões (seis categorias fundamentais) para cada classe de entidade.

Supondo as classes **rio**, **lago** e **estrada asfaltada,** tendo em conta as seis dimensões, três vetores (*hêxuplas* ordenadas) seriam formados por um método de natureza ontológica[251] de classificação:

- 1ª dimensão (substância) - valores: s_1 para água e s_2 para asfalto;
- 2ª dimensão (estado) - valores: e_1 para sólido, e_2 para líquido, e_3 para gasoso;
- 3ª dimensão (propriedade) – valores: p_1 para puntiforme, p_2 para linear, p_3 para poligonal ou planar e p_4 para forma complexa (combinação das outras);
- 4ª dimensão (operação) – valores: o_1 para suportar a substância, o_2 para substância em movimento;
- 5ª dimensão (reação) – não visualizada para os casos;
- 6ª dimensão (instrumento) – não visualizada para os casos.

É claro que poderiam ser testadas outras propriedades (objeto isolado x coleção, etc.) na terceira dimensão.

Combinando-se essa forma de definição por facetas com a expressão anterior de *estrutura formal da definição predicativa* (definição de "termo"), seria possível separar para cada termo os seguintes valores (foi retirado o componente "G", porque a taxinomia contemplá-lo-ia [V. Figuras 6.3(a, b, c)]:

rio ($\vec{r}$)= (s_1, e_2 , p_2, o_2)

lago ($\vec{l}$) = (s_1, e_2 , p_3, o_1)

estrada asfaltada ($\vec{e}$)= (s_1, e_2 , p_3, o_1)

É bom notar que os termos já estão no formato vetorial e que não foram usados todos os conceitos de categorização. Cada componente dos vetores é chamado de vetor unitário de um subespaço vetorial, no caso, 4-dimensional ou tertradimensional. Portanto, o conjunto das *4-uplas* ordenadas gera os subespaços em que estão aplicados os vetores $\vec{r}$ (rio), $\vec{l}$ (lago) e $\vec{e}$ (estrada asfaltada).

Essa tentativa seria muito interessante, mas ao mesmo tempo muito difícil de ser implementada e não será explorada na seção seguinte (o PRONTO©), que usará o *modus sciendi* da *estrutura formal da definição predicativa* (definição científica conotativa) para levantar as *fds* das entidades espaciais, sem usar um quadro referencial tão abstrato como o das categorias fundamentais.

---

[251] Extraídos notadamente de RICH (1993) e VICKERY (1975)

VICKERY (1975), que é da área das ciências da classificação, estabeleceu uma metodologia de classificação (análise por facetas) muito semelhante aos dos atuais autores da área da Ciência da Computação, especialmente GANESAN (2002), que recomenda, em primeiro lugar, **agrupar** o conteúdo informativo do domínio do saber por categorias (classificação por assunto) para, depois, classificá-las pela **ordem** (classificação taxinômica) ou pela hierarquia de generalização[252] e meronímica[253]), dentro de cada grande categoria.

Sem temer um excesso de simplificação, pode-se dizer que o MC da folha Faxinal foi montado seguindo essas diretrizes básicas.

As TBCD© é o resultado de um esforço de análise por facetas. Apesar de não explorar todo o potencial da definição científica, essas tabelas foram construídas com base nesses princípios da classificação: as definições *intensionais* que estabeleceram a grande malha de categorias da folha Faxinal e as situaram com mais precisão dentro dessa malha conceitual.

Relembrando, segundo FELBER (1984), as definições *por extensão* baseiam-se na enumeração de todas as instâncias (espécies) de uma classe (gênero) que estejam num mesmo nível de abstração.

Já as definições *intensionais* (predicativas ou conotativas) partem de um gênero imediato e das características que delimitam o conceito relativo a esse gênero, diferenciando-o de outros no mesmo nível de abstração. Nesta pesquisa, só há interesse nas definições *intensionais.*

Saindo do exemplo teórico das categorias fundamentais, na forma de matrizes de montagem de ontologias para as classes de entidades espaciais, foi preciso testar o PROFAX em bases mais práticas e de menor alcance conceitual, aplicando a Equação 3.1 a conjuntos de entidades (nós-folhas) enquadradas numa hierarquia (taxinomia) orlada para esse fim.

Todos os componentes de cada vetor são as folhas da árvore *n-ária* que está ilustrada nas Figuras 6.3 (a, b, c). Essa árvore possui seis níveis (0 a 5), tem altura (h) = 6, ou seja, da raiz até as folhas, contam-se cinco arestas (h -1 ou *internodalidade* da hierarquia).

Na Equação 3.1, como já descrito, c_1 e c_2 são as coleções de elementos (termos); o numerador do segundo termo da equação representa o produto interno entre dois vetores; o denominador, o produto das normas desses dois vetores (as coleções); w_1 e w_2 são valores normalmente associados a pesos arbitrários para cada elemento da coleção. No caso do exemplo de demonstração, todos os pesos foram considerados unitários para simplificar[254]

---

[252] Gênero-espécie (típicas relações hierárquicas).
[253] Todo-parte ou de agregação. Ë uma hierarquia anômala, cujas ressalvas serão expressas a seguir.
[254] É o princípio científico da "Navalha de Ockham" (RUSSEL, 1995) sendo aplicado.

os cálculos, já que a própria estrutura da árvore, como se verá, provê uma ponderação implícita[255] ao cálculo da SS.

A equação 3.1 pode ser interpretada geometricamente pelo ângulo que dois vetores fazem entre si, num espaço *n-dimensional*. Quanto maior o ângulo entre os vetores representativos das coleções de feições c_1 e c_2, menos similares são tais coleções[256].

A Figura 3.12, como já explicado, dá uma idéia da interpretação geométrica da SS.

A árvore das Figuras 6.3 (a, b, c) ordena os termos que denotam acidentes (entidades) do mundo real. Na raiz, encontra-se o nó ***entidade***. Dele, bifurcam-se duas subárvores: a da esquerda começa pelo nó **tangível** (entidades concretas); a da direita, que nem será objeto de investigação, pelo nó intangível (entidades abstratas). Daí em diante, com base no nó tangível, seguem-se uma nomenclatura e regras parecidas com as da modelagem que a 1ª DL adotou para a folha Faxinal (ferramenta da modelagem: UML™).

A árvore não é do tipo balanceada, i.e., não se ramifica de forma harmônica para formar subárvores simétricas em relação aos seus nós-raízes, porém, para efeito de harmonização genealógica, foram extirpados alguns nós interiores; p.ex: para a subcategoria ***infra-estrutura***, três subárvores foram geradas: ***transporte***, ***obra*** e ***energia***. De transporte, mais quatro seriam geradas (***rodoviário***, ***ferroviário***, ***hidroviário*** e ***dutoviário***). Foram eliminados esses quatro nós para manter uma harmonia de descendência com as entidades naturais que serão comparadas com as artificiais, normalizando a contagem de arestas ou braçadas (caminhos, trechos) dos nós-folhas até o nó-raiz (***entidade***).

Da ação de cortes anteriormente explanada, fica a questão: Estas simplificações e normalizações não podem afetar o resultado que o algoritmo de cálculo produzirá para a SS? A resposta é afirmativa, com base no que todos os autores que estudaram o fenômeno reportaram em seus experimentos. No entanto, esta afetação pode ser controlada (eliminá-la é um ideal inalcançável) por uma criteriosa definição dos termos e pela construção de uma estrutura hierárquica que reproduza, de maneira racional, as relações de generalização/especialização e de agregação/composição das entidades que "povoam" a realidade.

Os cálculos que a seguir serão sintetizados foram realizados segundo os teoremas e axiomas da Álgebra Linear e da Álgebra elementar. No entanto, não custa relembrar que o produto interno entre dois vetores de componentes idênticos terá como resultado a unidade; p.ex: se no produto de dois polinômios ocorrerem termos do gênero: $s_1 . s_1$ ou $o_2 . o_2$, os resultados serão iguais a 1.

---

[255] Esta ponderação vem da Lei de Tobler: "Coisas mais próximas são mais similares entre si".
[256] GANESAN (2002) denomina estas coleções de *bag of elements*.

Outro fato, já citado, é a ponderação implícita que a estrutura da árvore provê. Da raiz até as folhas, que são as feições ou componentes dos vetores, há um caminho marcado por cinco arestas. Portanto, a distância da subclasse hidrovia até o nó-raiz = 3/5, enquanto a de rio até o raiz = 5/5 = 1 (é a Lei de Tobler traduzida em números). Essas distâncias definem uma ordem (ponderação) de similaridade entre dois termos da estrutura com relação à classe que imediatamente os subordina [classe superordenada, que GANESAN (2002) chama de *lca* ou *lowest common ancestor*[257]]. No caso de cálculo de SS, é do *lca* que se contam as braçadas ou arestas da taxinomia.

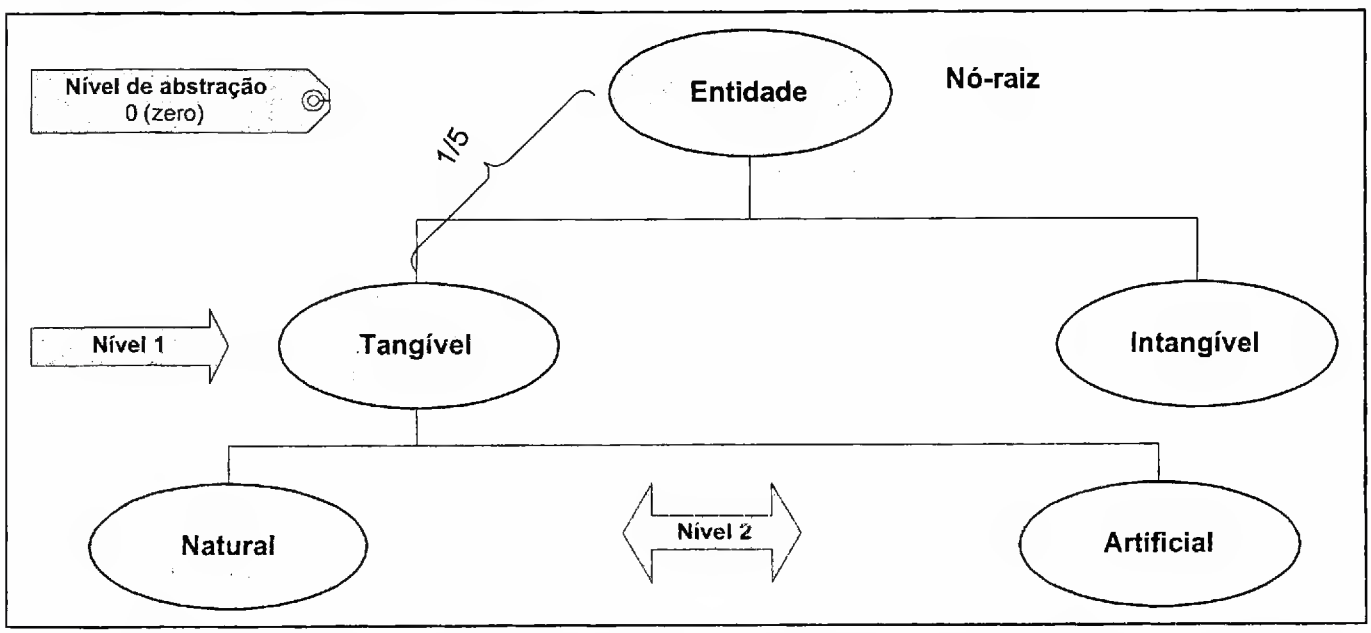

Figura 6.3(a): Árvore *n-ária* representativa da hierarquia que foi carregada no PROFAX.

Como já aventado, passa-se a um cálculo simples de similaridade entre dois pares de coleções (lembrete: cada coleção é um vetor). Volta-se a frisar que não se pode tratar cada termo por suas feições distintivas.

Coleção **A = {vala, rio}**. (classe hiperordenada ou *lca*: hidrografia)

Coleção **B = {rio, olho d'água}**. (*lca*: hidrografia)

Coleção **C = {rio, lago}**. (*lca*: hidrografia)

Coleção **D = {estrada de rodagem, ferrovia}**. (*lca*: transporte)

---

[257] RODRÍGUEZ(2000) denomina de *lub* (*least upper bound*), Jay J. Jiang [*apud* RODRÍGUEZ(2000}], também de *lub*, mas como *lowest upper bound*, P. Resnick [*apud* GANESAN (2002)], de *mub* (*minimal upper bound*)

É interessante notar o potencial da fórmula do co-seno, visto que o senso humano de similaridade entre esses pares de entidades fica bem representado pelos resultados numéricos.

Calculando a similaridade entre as coleções hídricas, i.e., entre A e B:

$$\vec{A} \bullet \vec{B} = (v,r) \times (r,o)$$

Este produto interno é resolvido pelo produto de dois binômios, como a seguir:

$$o.v + r.v + r.r + o.r$$

Mais uma vez lembrando: o produto interno de um vetor por ele mesmo = 1; os outros produtos terão seus resultados tirados do caminho de suas classes superordenadas até o primeiro nó comum entre elas. Para os outros dois produtos, esse caminho dá 3/5 para ***o . v*** e ***o . r*** e 4/5 para ***r . v***, sendo assim:

$$\vec{A} \bullet \vec{B} = \frac{3}{5} + \frac{4}{5} + 1 + \frac{3}{5} = 3{,}00$$

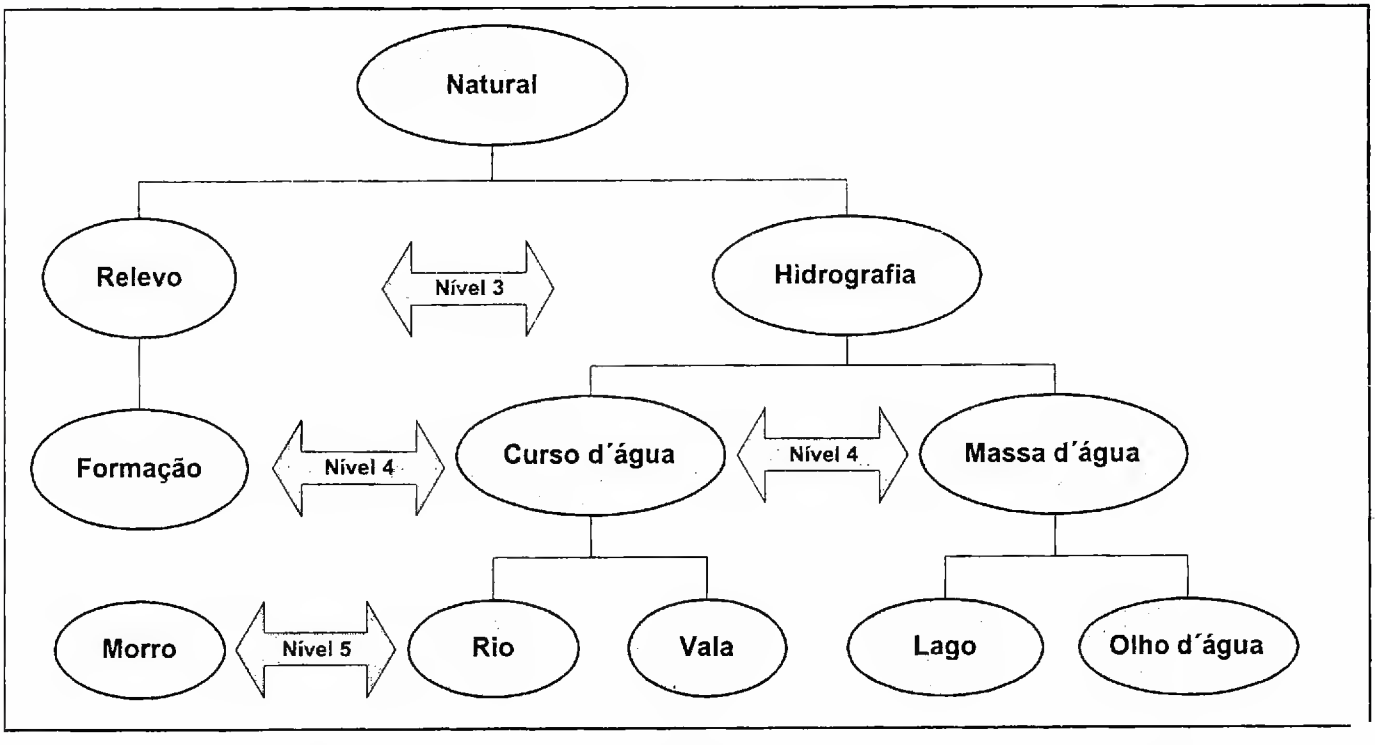

Figura 6.3(b): Continuação da árvore *n-ária* do PROFAX (categoria *Natural*).

Calculando o denominador, resultam como normas de A e B:

$$\left\|\vec{A}\right\| = \sqrt{(v \cdot v + r \cdot r + v \cdot r + r \cdot v)} = \sqrt{\left(\frac{18}{5}\right)}$$

$$\|\vec{B}\| = \sqrt{(r \cdot r + o \cdot o + r \cdot o + o \cdot r)} = \sqrt{\left(\frac{16}{5}\right)}$$

Juntando todos esses resultados parciais na fórmula da similaridade entre coleções hídricas, vem:

$$simil(\vec{A}, \vec{B}) = \frac{3{,}00}{\sqrt{\left(\frac{18}{5}\right)} \times \sqrt{\left(\frac{16}{5}\right)}} = 0{,}88$$

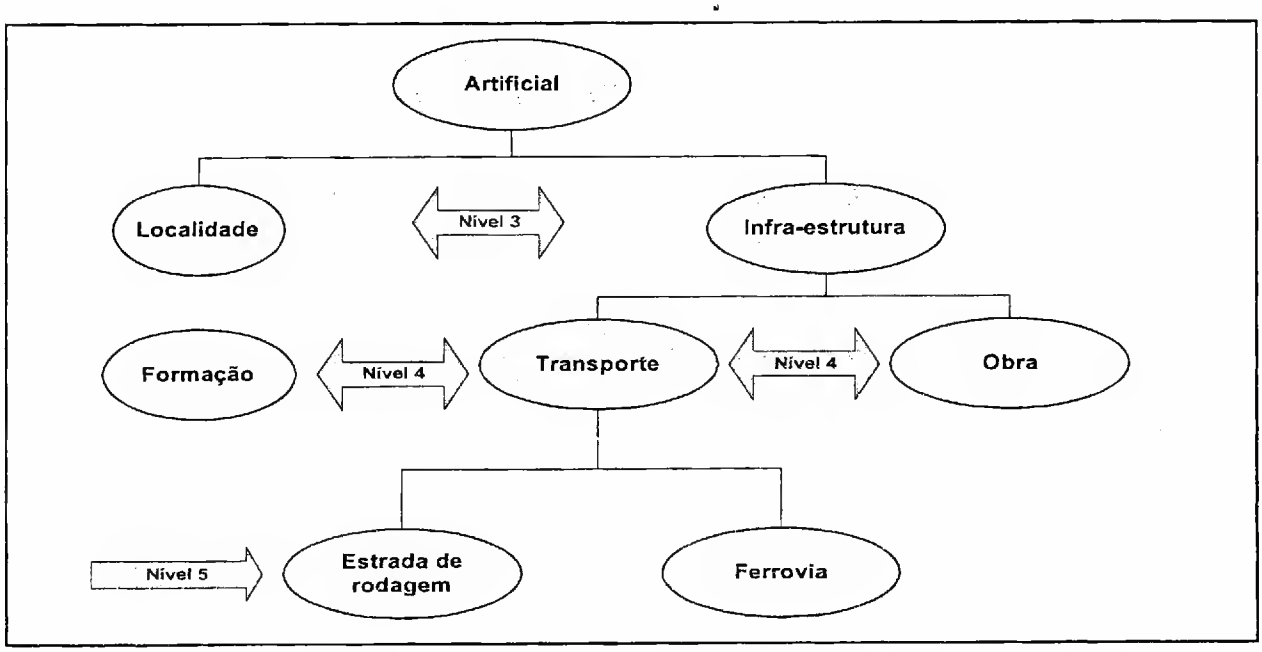

Figura 6.3(c): Continuação da árvore *n-ária* do PROFAX (categoria *Artificial*).

Resumir-se-á ainda mais a exibição dos cálculos da similaridade entre as coleções hídrica e de transporte. Nota-se que as classes superordenadas de curso d'água e de transporte ou de massa d'água e transporte estão bem longe umas das outras, próximas ao nó-raiz, numa distância de 1/5 deste nó.

Seja:

$$\vec{C} \bullet \vec{D} = (r, l) \times (e, f) = 0{,}80$$

$$\|\vec{C}\| = \|\vec{D}\| = \sqrt{\left(\frac{18}{5}\right)}$$

$$simil(\vec{C}, \vec{D}) = \frac{0{,}80}{\sqrt{\left(\frac{16}{5}\right)} \times \sqrt{\left(\frac{18}{5}\right)}} = 0{,}24$$

$\therefore simil(\vec{A},\vec{B}) > simil(\vec{C},\vec{D})$, confirmando a intuição humana.

Durante a demonstração, o PROFAX exibiu os mesmos resultados deste exemplo de treinamento. Os nós coloridos da árvore foram usados no cálculo de SS do exemplo.

A principal limitação para esse modelo de cálculo da similaridade semântica entre dois conjuntos de termos de uma dada língua profissional (cartográfica, no caso) é a impossibilidade de o modelo matemático (fórmula do co-seno) distinguir a SS entre os termos dos conjuntos selecionados de entidades, o que inviabiliza essa modalidade de avaliação para realizar a recuperação específica de objetos numa base de dados cartográficos.

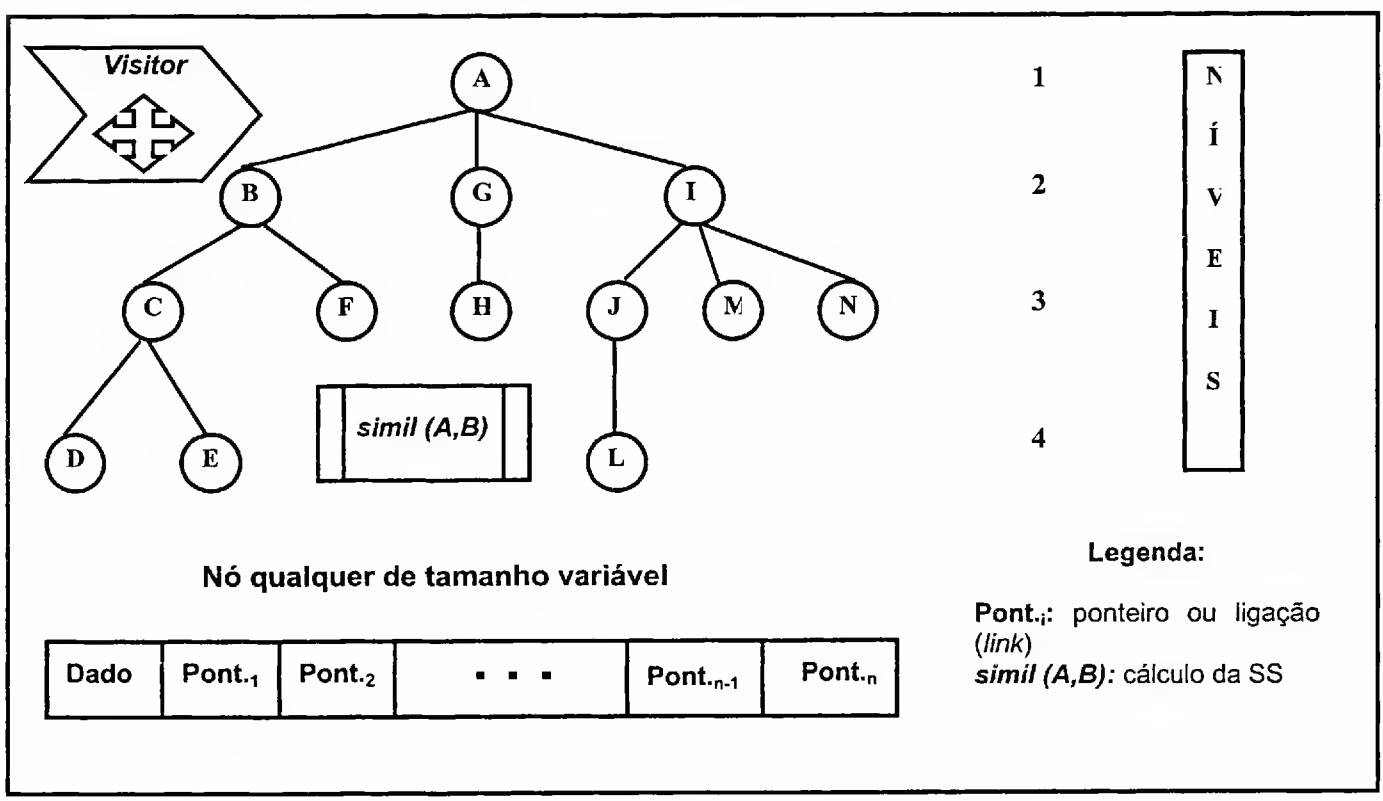

Figura 6.4: Primeira versão do PROFAX e arquitetura de um nó.

Apesar da limitação apontada, o objetivo dos autores que trabalham com esse modelo geométrico de SS não é o de pesquisar BDs cartográficos, mas bases de documentos textuais, em que um conjunto de termos (até frases) de busca do usuário deve ser comparado com uma base de documentos no formato digital, no intuito de orientar este usuário na descoberta de seu assunto de interesse nesta base.

O que se pretendeu com essa adaptação de problemas da área de busca de textos ao de busca de objetos em BDs foi o contacto com um modelo simples de cálculo de SS, que também já foi utilizado por pesquisadores de problemas de recuperação de informação geo-

gráfica e, também, para obter experiência na implementação de uma ferramenta automatizada de avaliação, como foi o caso do programa denominado PROFAX, de cujas mais de mil linhas de código em *Java*™, o que mais se acrescentou ao protótipo mais avançado que vem a seguir foi a construção de uma estrutura de dados flexível e bem adequada para organizar[258] os termos das entidades espaciais da folha Faxinal, como a da árvore *n-ária*.

A Figura 6.4 esquematiza a estrutura de dados escolhida para implementar a árvore *n-ária* - lista simplesmente encadeada (Figura 6.5). Esse foi o esquema da primeira versão para o PROFAX.

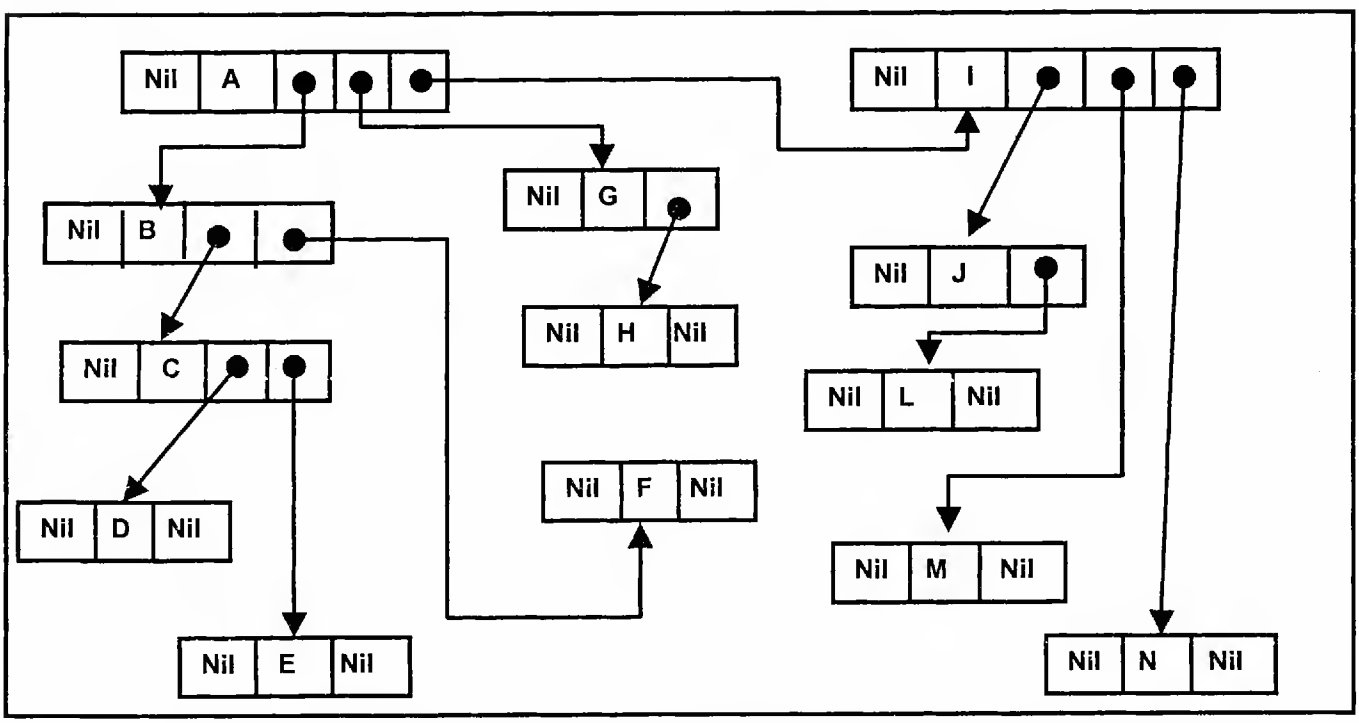

Figura 6.5: Lista simplesmente encadeada para a árvore *n-ária* do PROFAX.

A Figura 6.4 mostra ainda duas subestruturas importantes: a que contém a função de cálculo de SS e uma estrutura de dados (classe de usuário do *Java*™) denominada *Visitor*, cuja atribuição é percorrer (visitar) toda a estrutura da árvore *n-ária*.

O percurso realizado pelo *Visitor* é no sentido de cima para baixo e da esquerda para a direita (a ordem de distribuição das letras mostra isso), de acordo com as técnicas tradicionais de atravessamento de árvores [(*pré-ordem*, *pós-ordem* ou *na ordem* - HOROWITZ (1979)]. No caso do PROFAX, utilizou-se o atravessamento *pré-ordem*.

---

[258] Essa estrutura conforma-se à modalidade de representação do conhecimento da IA denominada rede semântica.

6.3.1.2. Extensão do modelo de avaliação de similaridade semântica: desenvolvimento do PRONTO© e construção de uma ontologia *ad-hoc*

Nessa nova versão do protótipo de avaliação de SS, foi abandonada a fórmula do co-seno entre dois vetores, uma vez que na versão anterior (PROFAX) só era possível comparar conjuntos de entidades (p.ex: rio+lago x ferrovia+rodovia), grupadas em vetores que respeitavam as métricas de um espaço euclidiano multidimensional, como foi demonstrado.

Naquela oportunidade, também foi explicado que a fórmula do produto interno, apesar de estimar a SS entre "pacotes" de entidades, era incapaz de verificar a SS entre os componentes de cada pacote e entre um componente de um certo pacote e seus ancestrais sucessivos, na estrutura hierárquica.

Nesta parte do trabalho, serão esquematizados os principais módulos do protótipo. Os esquemas seguintes guardam estreita ligação com os diagramas de classe e os diagramas de caso de uso do padrão UML™. Também serão descritas algumas características de desempenho do protótipo, suas limitações e sugestões de melhoria.

O presente exemplo de treinamento pode ser considerado como a especificação de requisitos que orientou a implementação do PRONTO©. Os três módulos do exemplo de treinamento estão nesta seqüência:

- Apresentação e explicação sumária das fórmulas do MSS [Modelo de Similaridade Semântica para uma ontologia isolada de RODRÍGUEZ(2000)];
- Montagem de quatro definições (RIO, LAGO, ESTRADA-DE-FERRO e CORPO D'ÁGUA CONTINENTAL) pela notação BNF, com base em RODRÍGUEZ (2002), nos documentos que foram compulsados do modelo conceitual do espaço geográfico brasileiro para a carta Faxinal (MCEGB - 1ª DL), do catálogo *on-line Wordnet*™, de livros e manuais técnicos de Geologia e Cartografia e da própria experiência do autor;
- Cálculo da similaridade semântica entre RIO e FERROVIA, entre RIO e LAGO, entre RIO e CORPO D'ÁGUA CONTINENTAL e entre CORPO D'ÁGUA CONTINENTAL e RIO.

A notação formal denominada BNF (*Backus-Naur Form*) é uma linguagem de representação adequada para a LPO (RUSSEL, 1995), que servirá de guia para a carga das ontologias no PRONTO©.

Esse protótipo, internamente, utiliza um modelo do padrão das linguagens de marcação (*markup languages*), denominado DTD[259] (V. glossário), para auxiliar o usuário na constru-

---

[259] Definição do Tipo de Documento [(padrão SGML, ISO 8879-1986 (ALMEIDA, 1999)].

ção da ontologia. Essa construção seria quase inviável por meios mais relaxados (menos algorítmicos), que permitissem mais intervenção humana.

De um esforço de extensão na revisão de literatura, é bom citar SMEATON (1997), que denominou essa estrutura de árvore *n-ária* implementada no PRONTO© *de grafo conceitual hierarquizado* (GCH ou HCG, *hierarchical concept graph*, no original).

No terceiro módulo, citado *ut retro*, os dois primeiros exemplos (rio *vs.* ferrovia e rio *vs.* lago) servem para comprovar o poder discriminatório da avaliação da SS, i.e., calculam-se os valores dessa variável para entidades designadas pelos termos cartográficos correspondentes, o que não era possível na versão do PROFAX. Os dois últimos exemplos testam o efeito da assimetria, i.e., verificam o deslocamento da carga de similaridade do **sujeito** (variante ou alvo) para o **predicado** (referente ou protótipo), provando que o valor da similaridade é maior na direção de um nó-filho para o nó-pai do que o contrário.

Evidentemente, a abstração coerente com a realidade na montagem da estrutura hierárquica e a construção de definições *intensionais* (conotativas) adequadas de cada termo, são fatores essenciais para que os resultados se mostrem satisfatórios por ocasião das comparações que se farão entre os resultados apurados no PRONTO© com as respostas dos questionários aplicados aos indivíduos do CCAuEx e da DSG.

A seguir, o primeiro módulo do exemplo de treinamento (**formulações**).

A Equação 6.1 é chamada de *modelo da razão ou normalizado de SS*, porque os valores que dela se obtêm variam entre [0, 1].

$$S_t (c_1, c_2) = \frac{[C_1 \cap C_2]_\#}{\{ [C_1 \cap C_2]_\# + \alpha[C_1 - C_2]_\# + (1 - \alpha)[C_2 - C_1]_\# \}} \quad \text{Eq. 6.1}$$

Em breves palavras, a fórmula expressa o cálculo da SS entre duas classes de entidades espaciais (c_1 e c_2), representadas pelos conjuntos de suas feições distintivas (*fds*) C_1 e C_2. Sua base é a Teoria dos Conjuntos e a Teoria de SS de Amos Tversky, cujos princípios gerais já foram vistos na revisão de literatura.

Recordando, um dos compromissos implícitos no objetivo geral deste trabalho está materializado nesta nova versão do protótipo, que processa as *fds* para poder avaliar a SS de forma individualizada entre as classes de entidades. Para isso, cada classe é decomposta nessas *fds* (*partes* (p), *funções* (f) e *atributos* (a) e o índice ***t*** de ***S*** (similaridade), nas Equações 6.1 e 6.2, significa o tipo de cálculo de similaridade para cada *fd*; quer dizer, a fór-

mula será empregada para cada *fd* e consolidada na similaridade total entre as duas classes pela Equação 6.2:

$$S_t (c_1, c_2) = w_p . S_p (c_1, c_2) + w_f . S_f (c_1, c_2) + w_a . S_f (c_1, c_2) \quad \text{Eq. 6.2}$$

A similaridade ***S***, portanto, é a soma ponderada das parcelas de similaridade (S_p, S_f e S_a), para cada *fd*.

Os pesos w_p, w_f e w_a podem ser obtidos por métodos complexos de avaliação do contexto, que fogem ao escopo deste trabalho, mas é possível considerar um contexto homogêneo para esse experimento, especialmente pelas condições ambientais (Sistema de Informações Geográficas) e o público-alvo escolhido, em que cada indivíduo possui uma formação básica similar e domina a sua língua profissional com propriedade, porquanto são todos profissionais de produção de documentos cartográficos. Sendo assim, diante desse fator de homogeneidade, a literatura [(RODRÍGUEZ, 2000) e RODRÍGUEZ, 2002]] autoriza atribuir um peso constante para cada coeficiente, que devem somar 1,00, i.e., w_p, = w_f = w_a = 0,333...

O que importa na Equação 6.1 é o seguinte:

- **c_1 e c_2:** classes de entidades, em que c_1 é o sujeito e c_2 é o referente; ambos são designados pelos seus termos cartográficos;
- ***t:*** tipo de *fd*: parte **(p)**, função **(f)** ou atributo **(a)**;
- **C_1 e C_2:** conjunto das feições do tipo ***t*** para as classes c_1 e c_2;
- **$C_1 \cap C_2$:** feições comuns a C_1 e C_2, i.e., que pertencem a C_1 e C_2;
- **$C_1 - C_2$:** feições que pertencem a C_1, mas não pertencem a C_2;
- **$C_2 - C_1$:** feições que pertencem a C_2, mas não pertencem a C_1.

Por questões de simplicidade, passa-se a chamar o termo **$C_1 \cap C_2$** de **X**, **$C_1 - C_2$** de **Y** e **$C_2 - C_1$** de **Z**.

A seguir, passa-se a explicar outro elemento importante dessas fórmulas: o coeficiente α, que vem expresso no denominador da Equação 6.1.

Esse coeficiente está ligado ao mesmo objetivo que a fórmula do co-seno cumpria no PROFAX, que era o de ajustar a similaridade entre dois termos (classes) pela distância semântica entre eles, na estrutura hierárquica da árvore *n-ária* (GCH). Por enquanto, só basta esta explicação e, na seqüência, verifica-se como ele se comporta nos exemplos.

As fórmulas de cálculo de α são as que vêm na Equação 6.3:

$$\alpha_{st}(c_1, c_2) = \begin{cases} \dfrac{d(c_1, l.u.b)}{d(c_1, c_2)} & \text{(a)} \quad | d(c_1, l.u.b) \leq d(c_2, l.u.b) \\ 1 - \dfrac{d(c_1, l.u.b)}{d(c_1, c_2)} & \text{(b)} \quad | d(c_1, l.u.b) > d(c_2, l.u.b) \end{cases}$$

Eq. 6.3

Em que:

- **$d(c_1, lub)$** é a distância entre a classe c_1 e o ancestral comum e imediato (*least upper bound – lub*) entre c_1 e c_2.

A fórmula de α fica mais fácil de compreender com um exemplo: Seja a estrutura hierárquica abaixo:

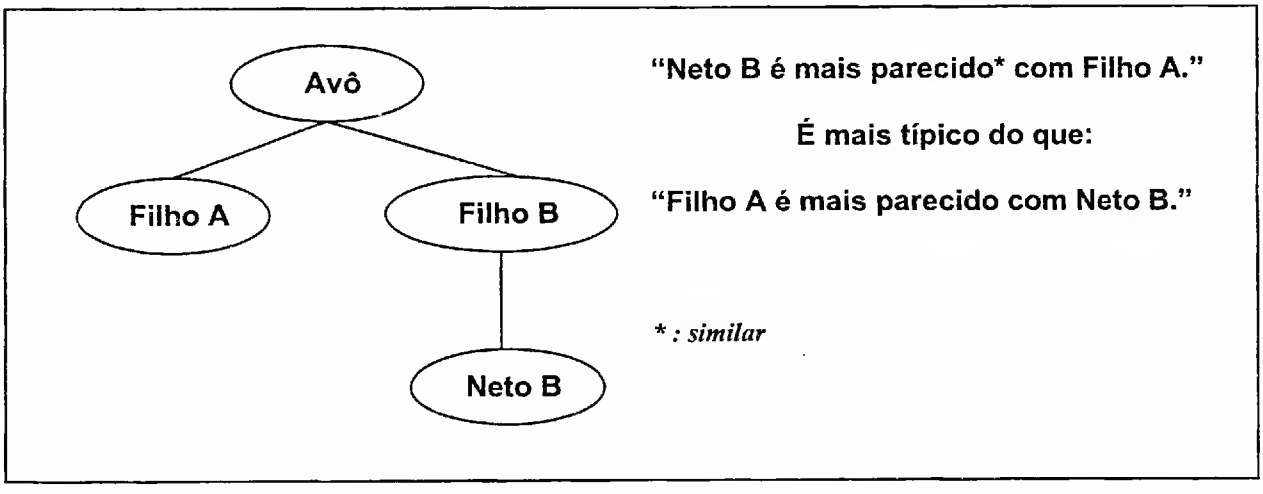

Figura 6.6: Efeito de assimetria numa taxinomia.

Como se vê na Figura 6.6., há três arcos (relações), ligando Neto B - Filho B – Avô – Filho A.

Pois bem, se o intuito é calcular a "distância semântica" entre Filho A e Neto B, este seria o resultado:

α (Filho A, Neto B) = 1/3 = 0,3...

Da mesma forma, se o intuito fosse a "distância semântica" entre Neto B e Filho A, viria:

α (Neto B, Filho A) = 1 – 1/3 = 2/3 = 0,67

Uma pergunta natural que surge é a seguinte: Como foi aplicada a Equação 6.3 e o que os resultados significam?

Percebe-se que o nível mais alto da hierarquia é o da classe Avô e o que está mais profundo é o da classe Neto B. A Equação 6.3 manda calcular o coeficiente α pela modalidade (a), se a distância do sujeito ao *lub* é menor ou igual à distância do referente ao *lub*

Sendo assim, α (Filho A, Neto B) corresponde a esta situação, porque a distância de Filho A para Avô (ancestral comum de Filho A e Neto B) é igual a 1/3 (0,3...) e a distância de Neto B para Avô é 2/3. Mudando de posição (Neto B, Filho A), a situação fica satisfeita pela outra fórmula do complemento de 1 (α = 0,67).

Quadro 6.1: Estrutura para definição de classes pela notação BNF.

<classe_de_entidades> ::= classe_de_entidades {

nome: {<*con_sin*>}
descrição: <*"descrição"*>
é-um(a): <é-um(a)>*
parte-de: <parte_de>*
todo-de: <todo_de>*
partes: <partes>
funções: <funções>
atributos: <atributos>**}**

==========Definições Formais==========

<descrição> ::= <palavra> | <*"descrição"*>; <palavra>
<é-um(a)> ::= { } | {<classe_entidades_pont>}
<parte-de> ::= { } | {<classe_entidades_pont>}
<todo-de> ::= { } | {<classe_entidades_pont>}
<partes> ::= { } | {<*con_sin*>}
<funções> ::= { } | {<*con_sin*>}
<atributos> ::= { } | {<*con_sin*>}
***<con_sins>* ::=** {<*con_sin*>} | <*con_sins*>; {*con_sin*}
***<con_sin>* ::=** <palavra> | <*con_sin*>; <palavra>
***<ponteiro_ para_classe_entidades>* ::=** <ponteiro> | <pt_classe_entidades>, <ponteiro>

Logo, o valor de α é um indicador do efeito da assimetria, que associa o sujeito e o predicado ao gradiente de similaridade, maior do mais específico para o mais genérico e vice-versa. É mais intuitivo dizer que um neto é parecido com seu tio (0,67), do que dizer

que um tio é parecido com o seu sobrinho (0,3...), adaptando o caso para uma situação quotidiana ou do senso comum.

A seguir, o segundo módulo do exemplo de treinamento (**as definições**).

O símbolo **{<*con_sin*>}**, na tabela *ut retro*, significa um conjunto de sinônimos de um determinado termo. Sua acepção se deve a um dos idealizadores da lexicografia *Wordnet*™, George A. Miller [*apud* RODRÍGUEZ (2000)], que os denominava de ***synset*** (*synonym set*). Aqui, adaptou-se a abreviatura *synset* para *consin*.

Os elementos encimados por um asterisco (*) representam classes vinculadas a outras classes (*ponteiro_para_classe_entidades*, na área de definição da Quadro 6.1).

A seguir (Quadro 6.2), um caso de definição para a classe **ESTÁDIO**:

Quadro 6.2: Notação BNF para a classe ESTÁDIO.

classe_de_entidades {

nome:	{ **estádio**, boliche, arena} }
descrição:	*"Ampla estrutura, geralmente descoberta, na qual se realizam competições desportivas."*
é-um(a):	{ {construção*} }
parte-de:	{ }
todo-de:	{ {campo_de_atletismo*} }
partes:	**{** {campo_de_atletismo, campo_desportivo, área_de_lazer}, {vestiário}, {estabelecimento}, {parte_central_de_um_campo_desportivo}, {arquibancada}, {posto_de_venda_de_ingressos, bilheteria} **}**
funções:	**{** {jogar, competir}, {brincar, praticar}, {divertir-se, tocar_um_instrumento_musical} **}**
atributos:	**{** {propriedade_arquitetônica}, {coberto/descoberto}, {nome}, {com_iluminação/sem_iluminação}, {privado/público}, {tipos_de_esportes}, {tipos_de_usuários} **} }**

Depois de montadas todas as definições das classes de entidades espaciais, estará construída a ontologia *ad-hoc*[260], sobre a qual o PRONTO© realizará a avaliação da SS. Ainda será preciso traduzi-la para uma notação mais rigorosa e normalizada dentro do protótipo, que é a DTD.

É necessário enfatizar um procedimento metodológico que difere para os dois campos de observação (V. Figura 6.1). No campo das definições para a montagem da ontologia *ad-hoc* (1° campo de observação), as definições das entidades não podem ser exatamente as mesmas que foram distribuídas para os respondentes (2° campo de observação). As definições fornecidas no questionário aos indivíduos, apesar de terem sido extraídas de glossários ou de manuais técnicos, não obedecem (e nem deveriam obedecer) às métricas da modela-

gem conceitual da folha Faxinal, até mesmo para manter a necessária independência entre as unidades de observação pertencentes aos dois campos de observação.

Quadro 6.3: Notação BNF para a classe LAGO.

classe_de_entidades {	
nome:	{ {**lago**, laguna, lagoa} }
descrição:	*"Grande corpo d'água (geralmente doce) cercado de terras."*
é-um(a):	{corpo d'água continental*}
parte-de:	{ }
todo-de:	{ { angra}, {foz*}, {vau*} }
partes:	{ {água}, {leito}, {angra, enseada}, {foz, delta, estuário, desembocadura}, {vau, passagem}, {margem} }
funções:	{ {navegar}, {transportar}, {vadear}, {pescar}, {entreter} } *Vadear: passar a vau, pelo lugar mais raso do rio.*
atributos:	{ {profundidade}, {forma}, {estado_físico}, {salinidade}, {extensão}, {navegabilidade}, {ilha}, {rocha}, {leito}, {banco_areia} } }

Sendo assim, a definição de ESTÁDIO que está sendo utilizada neste exemplo de treinamento (Quadro 6.2) não será a mesma que será usada para carregar a ontologia *ad-hoc* no PRONTO©.

Quadro 6.4: Notação BNF para a classe RIO.

classe_de_entidades {	
nome:	{ {**rio**, curso d'água, riacho, regato, córrego} }
descrição:	*"Corpo d'água, de largura e extensão variáveis, que escoa suas águas em virtude do desnível entre o local onde nasce e o local onde desemboca.."*
é-um(a):	{ {corpo d'água continental*} }
parte-de:	{ }
todo-de:	{ {foz*}, {vau*}, {canal*}, {cachoeira*}, {corredeira*}, {margem*}, {sumidouro*}, {olho d'água*}, {meandro*}, {linha_drenagem*}, {área_drenagem*} }
partes:	{ {água}, {leito}, {foz, delta, estuário, desembocadura}, {vau, passagem}, {canal, vala}, {margem, ribeira}, {sumidouro}, {olho d'água}, {cabeceira}, {eixo} } *Eixo: local do rio em que a sua velocidade de escoamento (vazão) é máxima.*
funções:	{ {navegar}, {transportar}, {vadear}, {pescar}, {entreter}, {represar} }
atributos:	{ {profundidade}, {vazão}, {forma}, {estado_físico}, {salinidade}, {extensão}, {navegabilidade}, {ilha}, {leito}, {rocha}, {praia}, {banco_areia} } }

Feita essa ressalva, seguem-se as tabelas de definições das classes LAGO (Quadro 6.3), RIO (Quadro 6.4), FERROVIA (Quadro 6.5) e CORPO D'ÁGUA CONTINENTAL (Quadro 6.6).

---

[260] Conjunto de todas as tabelas de notações BNF para as classes.

Os conjuntos de sinônimos (*consins*) entre chaves devem ser comparados com o homólogo do par da seguinte forma (***Axioma 1***): pontua positivamente todo o conjunto como uma unidade se no *consin* da outra entidade houver pelo menos um dos sinônimos coincidindo.

Alguns atributos como: {profundidade}, {forma}, {estado_físico}, {salinidade}, {extensão}, {navegabilidade}, neste exemplo de treinamento calculado à mão, foram comparados literalmente e não pelos valores (a maioria dicotômicos[261]) que podem assumir.

Para o caso de FERRROVIA e CORPO D'ÁGUA CONTINENTAL, foi elaborada uma definição expedita, que mesmo assim não comprometeu o resultado final do exemplo.

Quadro 6.5: Notação BNF para a classe FERROVIA.

classe_de_entidades {
nome: { {**ferrovia**} }
descrição: " - "
é-um(a): { {meio_transporte*} }
parte-de: { }
todo-de: { }
partes: { {dormente}, {trilho, carril}, {cravo}, {leito} }
funções: { {transportar, carregar} }
atributos: { {bitola}, {estação}, {túnel} } }

Este trabalho segue a linha de representação lexicográfica *um vocábulo – um termo* para as *fds*, em vez da puramente semântica, que admite termos compostos (mais de um conceito representado pelo termo) e até frases.

Quadro 6.6: Notação BNF para a classe CORPO D´ÁGUA CONTINENTAL.

classe_de_entidades {
nome: { {**corpo d'água continental**} }
descrição: *"Porção de um continente coberta de água. "*
é-um(a): {corpo d'água*}
parte-de: { }
todo-de: { {curso d'água*}, {lago*} }
partes: { {água}, {foz, delta}, {costa}, {ecossistema} }
funções: { {navegar}, {transportar, carregar}, {pescar}, {manter_vida} }
atributos: { {profundidade}, {estado_físico}, {impacto_ambiental}, {maré}, {recife} } }

Assim, uma *fd* é representada por um conjunto de sinônimos (*consins*) entre chaves e o algoritmo de comparação dos *consins* deve empregar uma operação exaustiva de verificação de coincidências das cadeias de caracteres que se referem a cada feição expressa pelo *consin*.

[261] Fundo/raso, veloz/lento, salgado/doce, etc.

Como dito, não são considerados termos (conceitos) compostos, mas vocábulos compostos, separados por símbolos de sublinhado ("_"), que representem uma classe de entidades espaciais, sem acrescentar problemas maiores ao resultado. Esta regra constitui um outro axioma de formação (***Axioma 2***)[262], que será seguido pela da ontologia DTD (na forma de um arquivo XML™).

A vantagem desse processo de verificação exaustiva da coincidência entre cadeias de caracteres entre *consins* é poder varrer rapidamente uma ontologia muito carregada de definições. Essa forma de verificação é chamada por RODRÍGUEZ (2002) de *similaridade semântica restritiva*.

Assim, a nova versão do protótipo deve comparar os conjuntos de sinônimos, levantar as coincidências entre os vocábulos e aplicar essas pontuações das coincidências nas fórmulas dadas. Essas pontuações (ou cardinalidade) são indicadas na Equação 6.1 pelo símbolo "#".

É interessante confrontar autores tão distantes no tempo e no domínio do conhecimento como RODRÍGUEZ (2000, p.40), da IA, e VICKERY (1975, p. 45), da Biblioteconomia. Ambos enfrentaram problemas semelhantes na construção de estruturas de representação do conhecimento, diferindo apenas na forma de solucionar os problemas. O segundo autor, sem dispor de processos automatizados de tratamento de dados, utilizou processos manuais e algoritmos de ordenação empíricos, com muita intervenção humana, o que tornou a metodologia menos formal, estanque e, por conseguinte, mais sujeita a erros. A primeira autora, já usufruindo de meios mais avançados de tratamento de dados, pode ter superado vários óbices instransponíveis para o segundo autor, mas não deixou de ser estorvada pela complexidade de ajustar ao ambiente rigidamente formal de um sistema computacional a plasticidade e o dinamismo das entidades e de seus comportamentos no MR.

Arranjar os termos de uma taxinomia numa seqüência útil é um problema tradicional de classificação. O primeiro ponto é decidir se todos os termos incluídos numa categoria podem ser arranjados numa única árvore hierárquica. Isso não será possível, se a área de estudo não for ainda consistente. É a chamada "classificação cruzada" pelos cientistas da classificação (VICKERY, 1975, p. 46) ou a "herança múltipla"[263] para os cientistas da computação (FURLAN, 1998, p. 304) ou as chamadas relações diagonais da Terminologia (FELBER, 1984).

---

[262] O Axioma 29 complementará este axioma, considerados, ambos, os mais relevantes da ontologia *ad-hoc*.
[263] Variação semântica da generalização, em que um tipo pode ter mais de um supertipo.

VICKERY (1975, p.46), da mesma forma que diversos autores contemporâneos pesquisados no campo da IA, assinala que a classificação cruzada é um sinal de que é necessário interferir na sistematização do conhecimento, particularmente numa área em que ainda não houve um acordo geral quanto às características das classes mais genéricas do domínio do conhecimento.

Na verdade, a injunção *ut supra* foi expressa por RODRÍGUEZ (2000, p.40) para prover o seu MSS com os axiomas de formação da hierarquia de uma rede semântica. Suas restrições axiomáticas foram implementadas na linguagem técnica de programação que utilizou para construir o seu protótipo - o C++™.

É bom frisar que a relação eminentemente hierárquica e lógica é a de generalização (hiperônimo – hipônimo). Porém, neste estudo-de-caso, as relações partitivas (holônimo – merônimo), de natureza ontológica, também foram consideradas hierárquicas, exatamente como propôs RODRÍGUEZ (2000), com a ressalva de que elas não contam com o mecanismo da herança, específico da de generalização. Essa restrição abranda algumas dificuldades de implementação, mas dá ensejo para tornar vulneráveis alguns princípios da OO.

A mesma dificuldade foi experimentada na elaboração do PRONTO© e as soluções adotadas foram as seguintes:

- Para o caso das relações meronímicas afetadas por hierarquia, o módulo de edição do protótipo permite derivar uma classe que seja parte ou componente de uma outra (holônimo). Isto constitui um outro axioma (***Axioma 3***) carregado na DTD;
- Para o caso da herança múltipla, o que se fez foi contornar o problema, eliminando-se esse tipo de relação dos diagramas de classes adaptados do modelo conceitual da folha Faxinal. Dessa forma, foram privilegiadas as relações essencialmente hierárquicas de gênero-espécie e as meronímicas adaptadas à hierarquia.

Deixando de lado as relações mais peculiares entre classes de entidades, que podem até indicar anomalia na estrutura de representação do conhecimento, passou-se para a apreciação daquelas relações mais privilegiadas pelos SIs em geral, como o PRONTO©, p.ex. FELBER (1984) classificou-as em dois subgrupos do grande grupo das relações verticais:

- Relações lógicas verticais
- Relações todo-parte

As primeiras são também conhecidas na OO como relações de generalização e de especialização (FURLAN, 1998). No âmbito da Lingüística Computacional e da Terminologia científica, são conhecidas como relações hiperonímicas e hiponímicas.

A segundas são também conhecidas na OO como relações de agregação ou de composição (FURLAN, 1998). No âmbito da Lingüística Computacional e da Terminologia científica, são conhecidas como relações holonímicas e meronímicas.

Ao iniciar o teste de desempenho do PRONTO©, calculou-se a SS nessa ordem: (rio, ferrovia) e (rio, lago). Depois: (rio, corpo d'água continental) e (corpo d'água continental, rio).

Nos primeiros dois pares, verifica-se a similaridade entre entidades individualizadas com um ancestral distante e um próximo, respectivamente.

Nos dois últimos pares, pretende-se testar o efeito da assimetria, já exemplificado no caso do avô, neto e tios.

Os *consins*, até certo grau, resolvem alguns problemas semânticos importantes, como o da ambigüidade[264], porque atribui por equivalência ao termo de interesse um conjunto de sinônimos que pode induzir um sentido menos polissêmico do termo, que a ontologia tem capacidade de expressar.

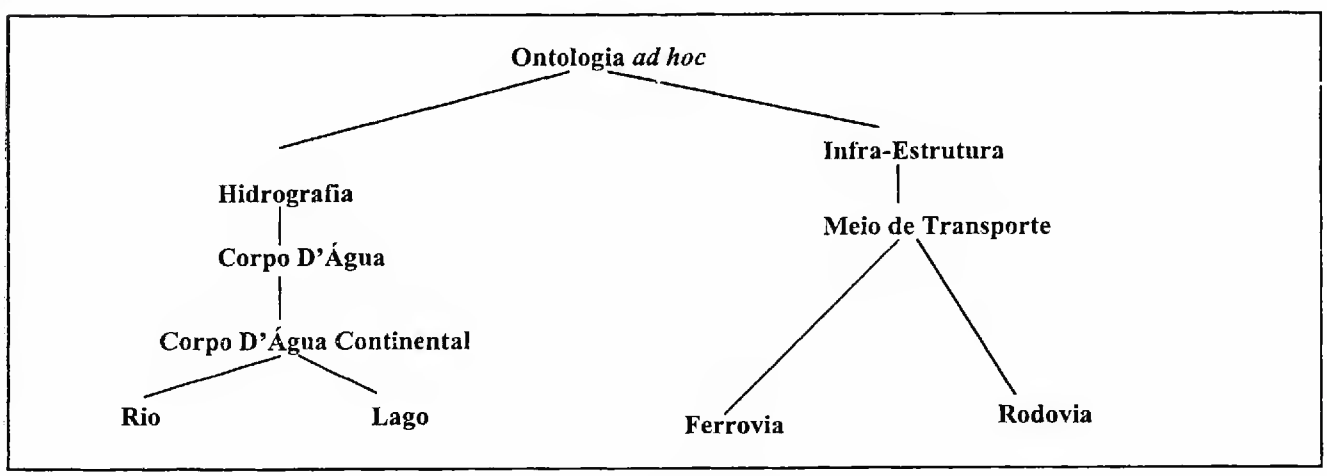

Figura 6.7: Taxinomia de treinamento para o PRONTO.

No GCH anterior, as relações de generalização foram discriminadas.

Passa-se, agora, para a terceira etapa: **os cálculos**.

O primeiro será para determinar a similaridade entre rio e ferrovia.

S (rio, ferrovia)

$|X| = rio.partes \cap ferrovia.partes =$ {leito} $\therefore \underline{1}$ (uma) coincidência.

[264] Não é o objetivo desse estudo. MEDEIROS (1999), no entanto, aprofunda a investigação desse assunto.

$|Y| = rio.partes - ferrovia.partes =$ { {água}, {foz, delta, estuário, desembocadura}, {vau, passagem}, {canal, vala}, {margem, ribeira}, {sumidouro}, {olho d'água}, {cabeceira}, {eixo} } ∴ 9 (nove) itens levantados.

$|Z| = ferrovia.partes - rio.partes =$ { {dormente}, {trilho, carril}, {cravo} } ∴ 3 (três) itens levantados.

Desta forma:

$$S_p(rio, ferrovia) = \frac{|X|}{(|X| + \alpha \times |Y| + (1-\alpha) \times |Z|)}$$

Da figura que ilustra a taxinomia das classes em apreço, fica fácil perceber que entre rio e ferrovia, a primeira classe está mais distante do ancestral comum (*lub*) de ambas, o que leva à aplicação da Equação 6.3 (b):

$$\alpha = 1 - \frac{4}{7} = \frac{3}{7}$$

Levando todos esses valores para a parcela da Equação 6.2, correspondente à similaridade entre as partes do par (rio, ferrovia), vem:

$$S_p(rio, ferrovia) = \frac{1}{\left(1 + \frac{3}{7} \times 9 + \frac{4}{7} \times 3\right)} = 0{,}15$$

Aplicando o mesmo procedimento para as parcelas das funções e dos atributos, vem:

$$S_f(rio, ferrovia) = \frac{1}{\left(1 + \frac{3}{7} \times 5 + \frac{4}{7} \times 1\right)} = 0{,}27$$

$$S_a(rio, ferrovia) = \frac{0}{\left(0 + \frac{3}{7} \times 12 + \frac{4}{7} \times 3\right)} = 0$$

Levando os resultados parciais das *fds* para a Equação 6.2, vem:

$$S(rio, ferrovia) = 0{,}\dot{3} \times 0{,}15 + 0{,}\dot{3} \times 0{,}27 + 0{,}\dot{3} \times 0 = 0{,}14$$

S (rio, ferrovia) = 0,14

Então, a similaridade entre rio e ferrovia, nas condições estipuladas pela ontologia aqui orlada, é de 0,14, numa escala que vai de zero a um.

S (rio, lago)

$$S_p(rio, lago) = \frac{4}{(4 + 0{,}5 \times 5 + 0{,}5 \times 1)} = 0{,}57$$

Aplicando o mesmo procedimento para as parcelas das funções e dos atributos, vem:

$$S_f(rio, lago) = \frac{5}{(5 + 0{,}5 \times 1 + 0{,}5 \times 0)} = 0{,}91$$

$$S_a(rio, lago) = \frac{10}{(10 + 0{,}5 \times 1 + 0{,}5 \times 0)} = 0{,}95$$

Levando os resultados parciais das *fds* para a Equação 6.2, vem:

$$S(rio, ferrovia) = 0{,}\dot{3} \times 0{,}57 + 0{,}\dot{3} \times 0{,}91 + 0{,}\dot{3} \times 0{,}95 = 0{,}81$$

S (rio, lago) = 0,81

Então, a similaridade entre rio e lago, nas condições estipuladas pela ontologia aqui orlada é de 0,81, numa escala que vai de zero a um.

S (rio, corpo d'água)

$$S_p(rio, corpo\_d'água) = \frac{2}{(2 + 0 \times 8 + 1 \times 2)} = 0{,}5$$

Aplicando o mesmo procedimento para as parcelas das funções e dos atributos, vem:

$$S_f(rio, corpo\_d'água) = \frac{3}{(3 + 0 \times 3 + 1 \times 1)} = 0{,}75$$

$$S_a(rio, corpo\_d'água) = \frac{2}{(2 + 0 \times 10 + 1 \times 3)} = 0{,}40$$

Levando os resultados parciais das *fds* para a Equação 6.2, vem:

$$S(rio, corpo\_d'água) = 0,\dot{3} \times 0,5 + 0,\dot{3} \times 0,75 + 0,\dot{3} \times 0,40 = 0,55$$

S (rio, corpo d'água) = 0,55

Então, a similaridade entre rio e corpo d'água, nas condições estipuladas pela ontologia aqui orlada é de 0,55, numa escala que vai de zero a um.

S (corpo d'água, rio)

$$S_p(corpo\_d'água, rio) = \frac{2}{(2 + 0 \times 2 + 1 \times 8)} = 0,20$$

Aplicando o mesmo procedimento para as parcelas das funções e dos atributos, vem:

$$S_f(corpo\_d'água, rio) = \frac{3}{(3 + 0 \times 1 + 1 \times 3)} = 0,50$$

$$S_a(corpo\_d'água, rio) = \frac{2}{(2 + 0 \times 3 + 1 \times 10)} = 0,17$$

Levando os resultados parciais das *fds* para a Equação 6.2, vem:

$$S(corpo\_d'água, rio) = 0,\dot{3} \times 0,20 + 0,\dot{3} \times 0,50 + 0,\dot{3} \times 0,17 = 0,29$$

S (corpo d'água, rio) = 0,29

Então, a similaridade entre corpo d'água e rio, nas condições estipuladas pela ontologia aqui orlada, é de 0,30, numa escala que vai de zero a um.

Mais uma vez, ficou demonstrado o efeito da assimetria, agora de forma mais circunstanciada, i.e., com fundamentação numa estrutura formal construída para avaliar a SS entre uma classe de entidades espaciais genérica - como *corpo d'água continental* -, e uma classe de entidades como *rio* - uma espécie de corpo d'água. Como 0,30 < 0,55, fica evidente que a carga de similaridade se desloca de *rio* na direção de *corpo d'água continental* e não o contrário.

Tendo sido carregado o exemplo de treinamento no PRONTO© e demonstrada a sua capacidade de calcular a SS para uma ontologia simples como a que foi ilustrada nos seus aspectos hierárquicos pela Figura 6.7, preenchida com *fds* oriundas das notações BNF das

Tabelas 6.3 a 6.6, chegou o momento de passar o protótipo por testes mais rigorosos. Para isso, uma parte do *corpus*, como já explanado várias vezes ao longo do texto, foi preparado um subconjunto do MC da folha Faxinal, acrescendo-se-lhe estruturas formais de definição predicativa para cada termo.

Esses testes basearam-se na bibliografia estatística voltada para estimativas não-paramétricas, como será explanado em mais minúcias nos próximos subitens.

Antes dos cálculos, é preciso expor e explicar os diagramas das classes de entidades espaciais da folha Faxinal que foram afetadas pelas avaliações dos indivíduos, nos questionários, e pelo PRONTO©, no seu procedimento de cálculo.

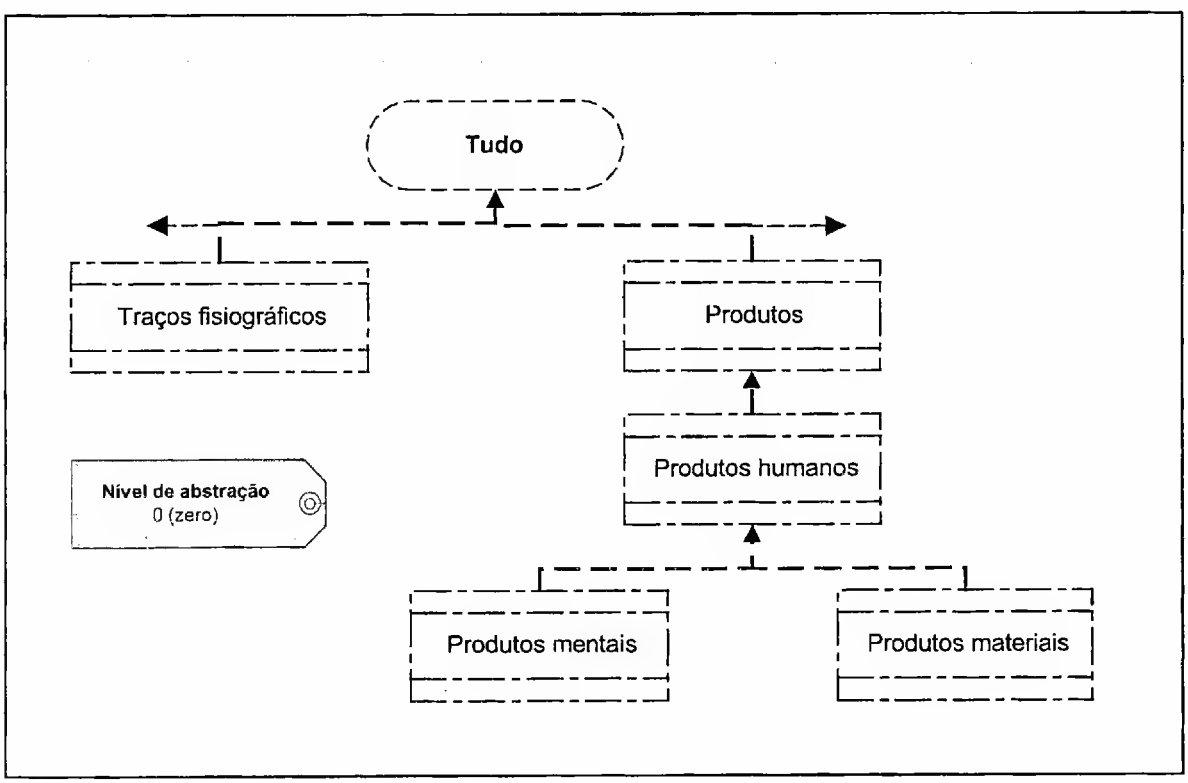

Figura 6.8: Categorias fundamentais do conhecimento cartográfico.

A Figura 6.8 é de um nível de abstração muito alto[265] e nem participou da ontologia da folha Faxinal[266]. A taxinomia nela retratada revela o processo tradicional de classificação descrito por VICKERY (1975), que mantém estreito relacionamento com os fundamentos conceituais da IA para construção de ontologias [(RICH, 1993), (RUSSELL, 1995) e RO-

[265] Ontologia genérica (V. Figura 3.15).
[266] Ontologia de domínio (*id.*).

DRÍGUEZ, 2000, *passim*]], pelo qual se levantam categorias fundamentais do conhecimento humano, capazes de abrigar as classes de entidades de níveis mais baixos e com maiores conteúdos informativos.

Para facilitar os cálculos de SS, essas grandes categorias do conhecimento (nível zero) foram excluídas da árvore *n-ária*.

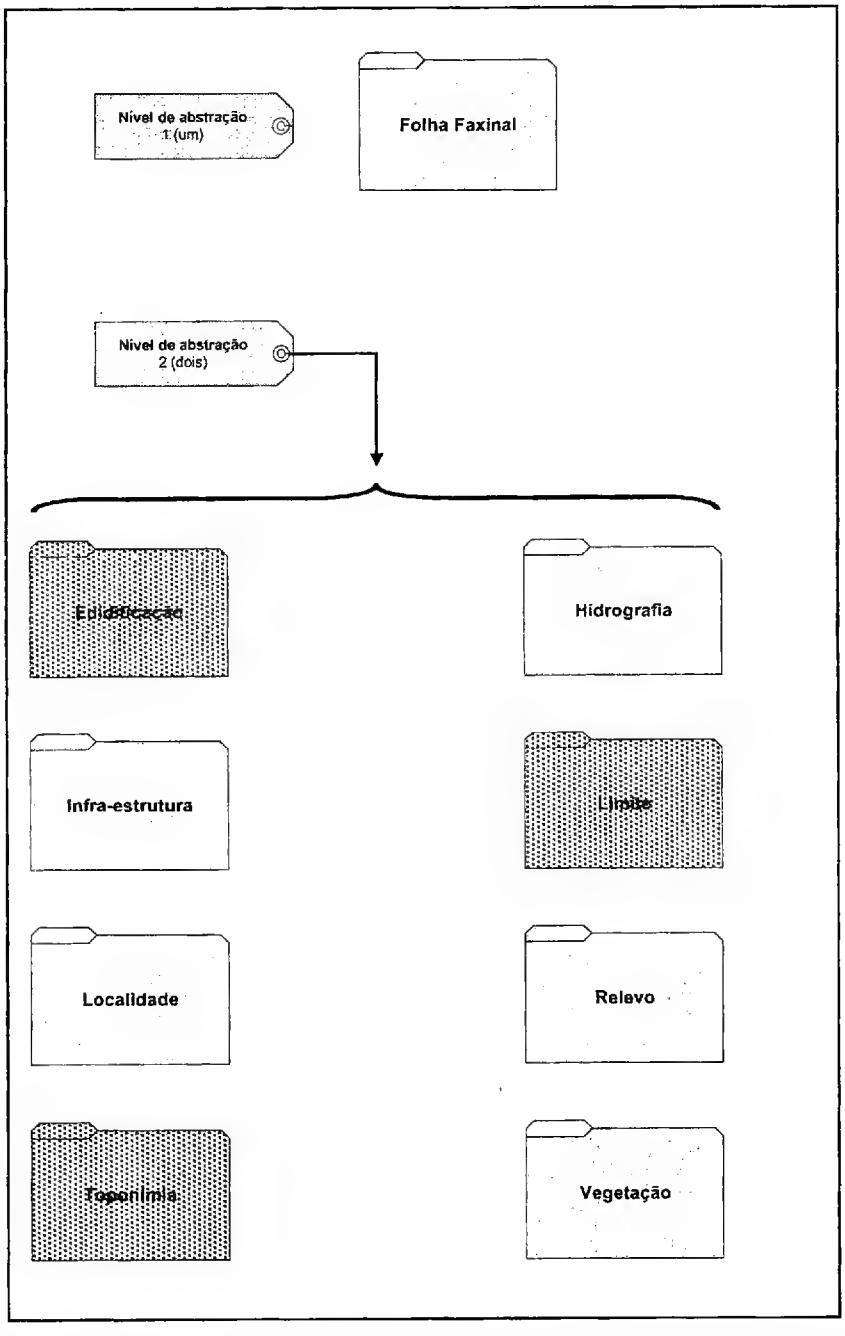

Figura 6.9: Categorias do projeto da carta topográfica da região de Faxinal (PR).

As perguntas básicas nessa fase de delineamento de uma ontologia são as seguintes:

- Quais são as feições (traços, características ou propriedades essenciais) dos termos dentro de uma categoria em relação aos de outras categorias?
- Quais são as feições (*fds* ou propriedades secundárias) que melhor definem um termo dentro de uma categoria?

Essas perguntas traduzem de forma interrogativa os esquemas de representação de uma definição científica expostos na *estrutura formal da definição predicativa* (V. p. 146, subitem 6.3.1.1 e Tabela 3.1).

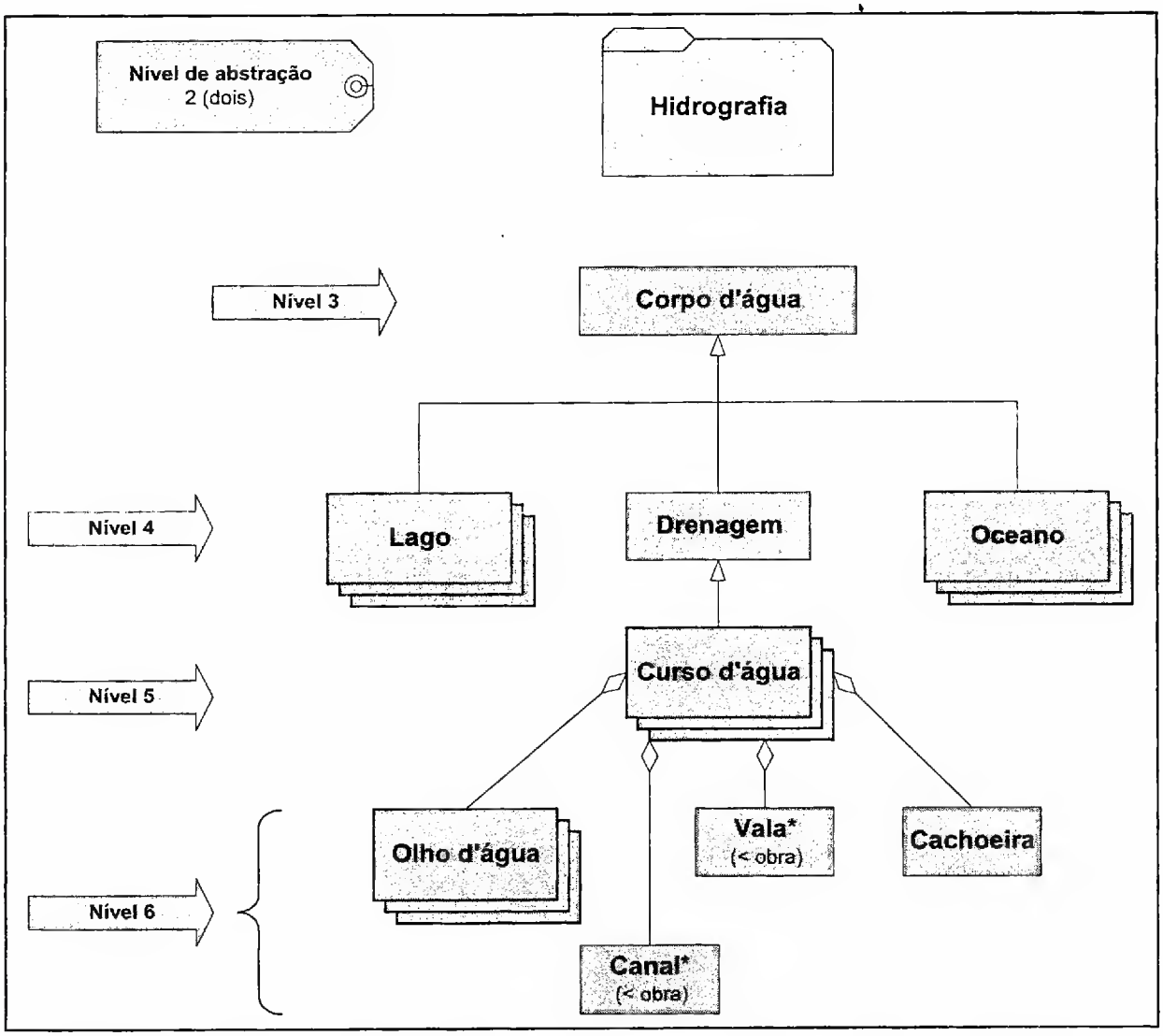

Figura 6.10: Relações de generalização e agregação para a categoria hidrografia.

A primeira pergunta sonda a natureza exógena das características do conceito, definindo-o e classificando-o em função da relevância dessas características ao longo de uma hierarquia. A segunda pergunta sonda a natureza endógena das propriedades do conceito,

descrevendo-o e diferenciando-o de outros dentro de uma mesma classe, em função da relevância das propriedades comparadas entre si.

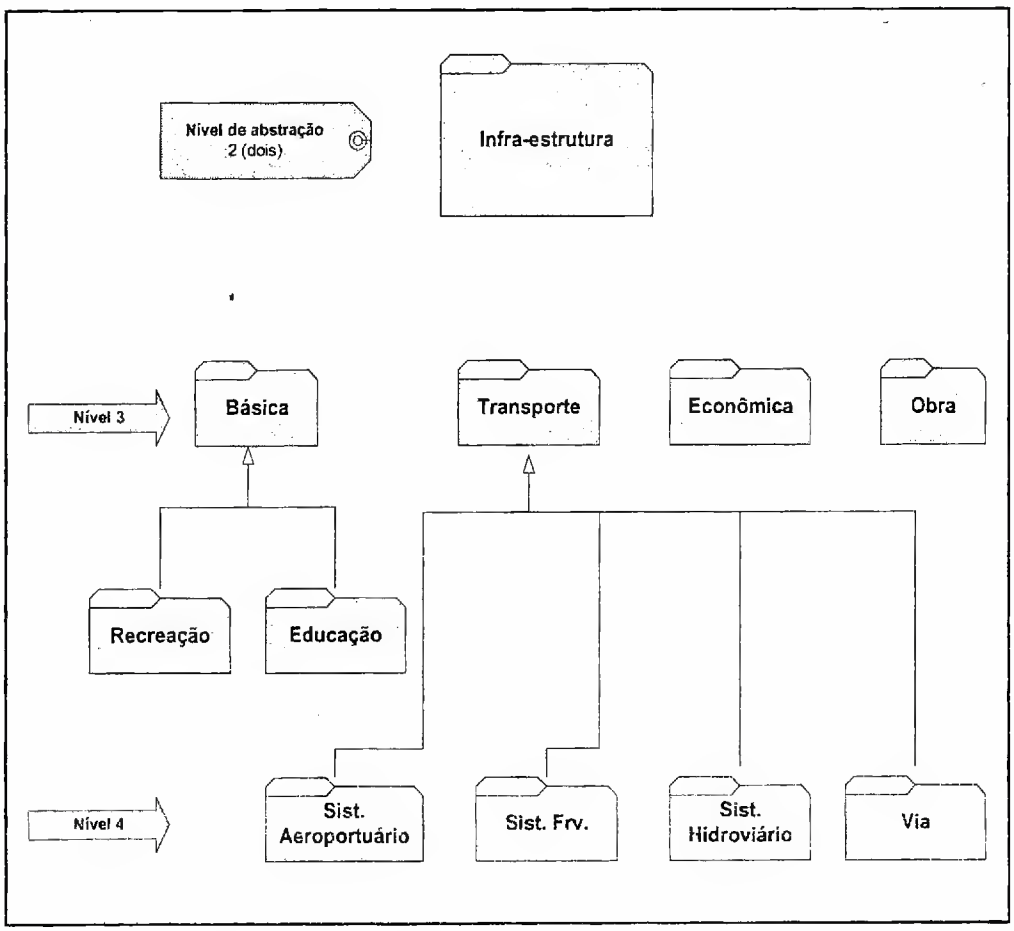

Figura 6.11: Subcategorias da categoria infra-estrutura.

Todas as notações gráficas dos diagramas de classe que vão da Figura 6.8 a 6.22 seguiram bem de perto as especificações da UML™, particularmente para a representação das classes e das relações entre as classes. Nesse caso, as relações de *gênero-espécie* (OO) ou *hiperonímicas-hiponímicas* (Lingüística Computacional) foram representadas por setas. A seta aponta para a superclasse (OO) ou classe superordenada (Lingüística Computacional). Na UML™, trata-se de um relacionamento típico de taxinomia (classificação, subentendendo hierarquia) entre uma coisa mais geral e uma específica, que herda parte das características da mais geral que a subordina (FURLAN, 1998, p.135), diferenciando-se daquela por um aporte extra de informações especializadas (propriedades). Um objeto da subclasse é um tipo, espécie ou instância de objeto de uma superclasse. São as relações lógicas de FELBER (1984).

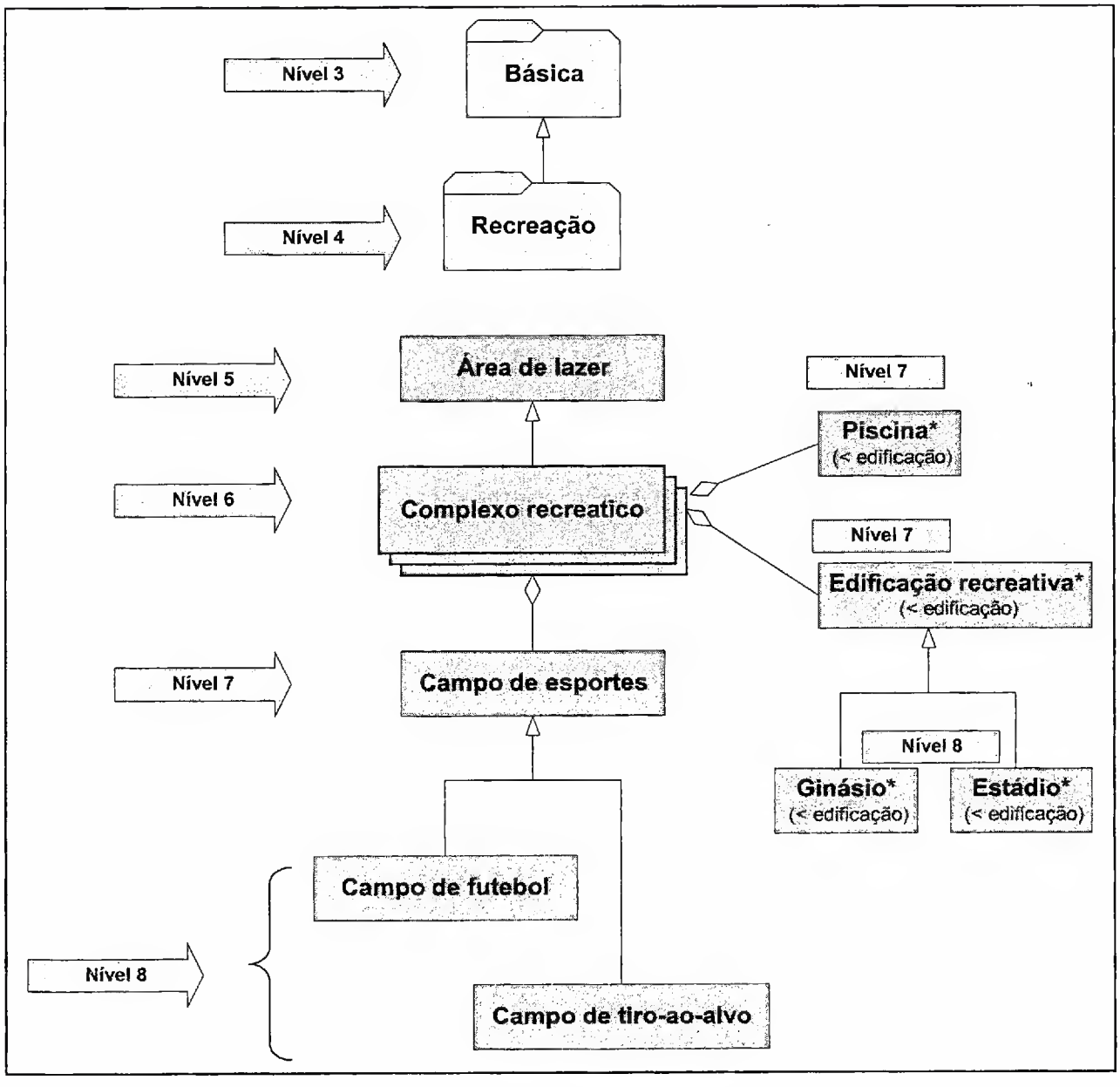

Figura 6.12 : Relações de generalização e agregação para a subcategoria recreação.

No caso das relações de *agregação* (OO) ou *holonímicas-meronímicas*[267] (ou ainda *to-do-parte* da Lingüística Computacional), a representação simplificada foi a de agregação regular ou por referência da UML™ (FURLAN, 1998, p. 129), i.e., um segmento de reta terminado por um losango ou diamante vazado, em que o diamante toca a classe holonímica (o todo). Essas relações mostram que um tipo de objeto é composto, pelo menos em parte, por outro(s) objeto(s). São as relações ontológicas de FELBER (*1984*). Semanticamente, indi-

[267] Segundo M. Winston *et al.* [*apud* RODRÍGUEZ (2000, p. 49)], há seis tipos: **componente-objeto**, membro-coleção, porção-massa, coisa-objeto, feição-atividade e lugar-área, dais quais só a em negrito foi considerada no PRONTO©.

cam que o objeto-parte é um atributo do objeto-todo e que a vida ou existência do primeiro, em geral, depende da existência do segundo.

No MC da folha Faxinal ainda há outros tipos de relação. Uma das mais freqüentes é a de *associação*, que denota relacionamentos entre classes não correlatas (corpo d'água e pântano, p.ex.); ou um relacionamento que descreve um conjunto de vínculos (conexões semânticas entre tipos de objetos); ou ainda para representar uma dependência estrutural entre objetos (corpo d'água e banco de areia, p.ex.).

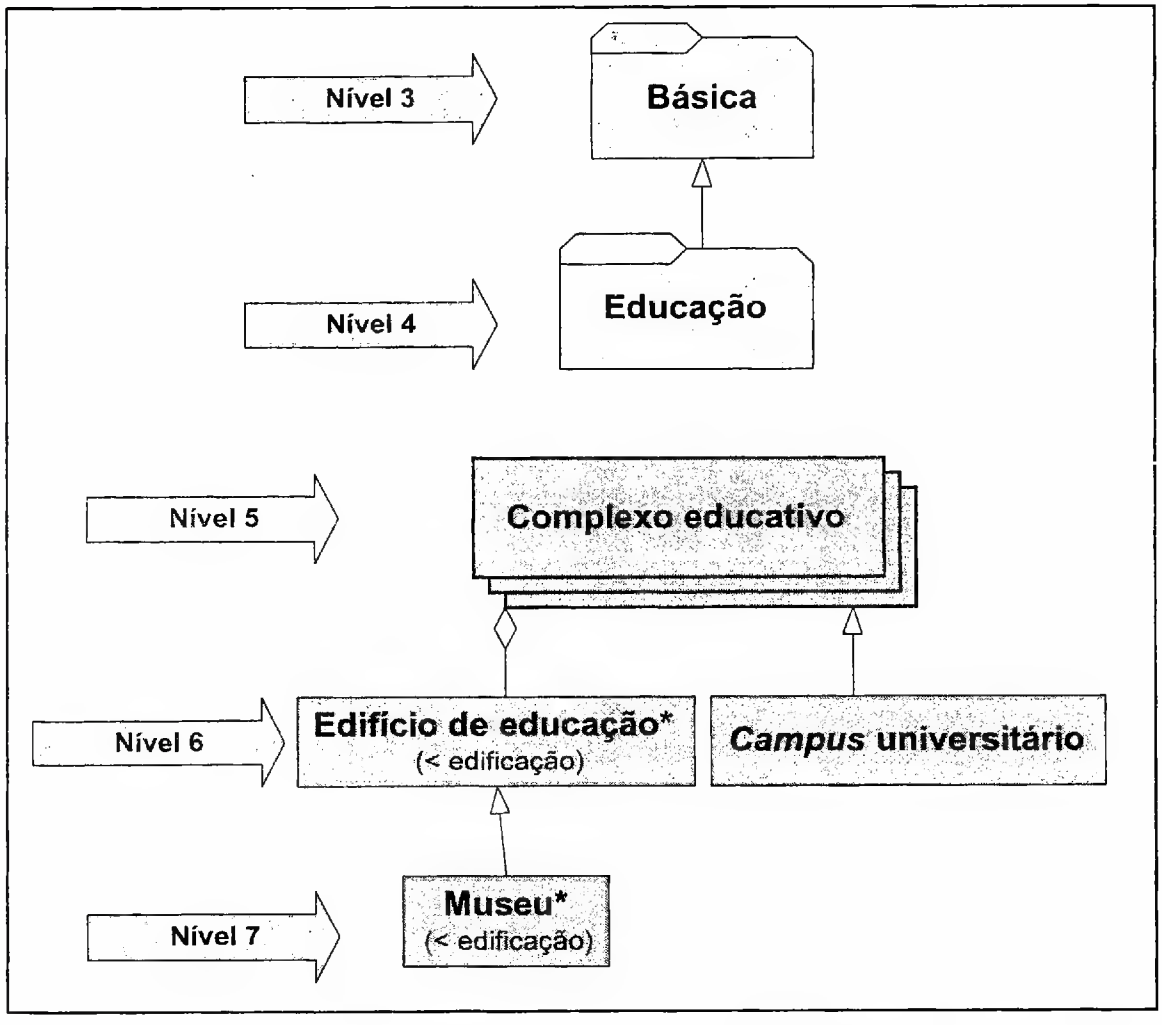

Figura 6.13 : Relações de generalização e agregação para a subcategoria educação.

A fim de coadunar a complexidade do MC da folha Faxinal às diversas definições de entidades geográficas extraídas de glossários, manuais, livros especializados e lexicografias *on-line*, como a *Wordnet*™, as relações de associação foram transformadas ou em relações de gênero-espécie ou em relações de agregação.

A única alteração à notação da UML™ ficou por conta dos nomes das classes. A regra de notação foi desenhar uma seqüência de retângulos quando um termo possuir mais de uma denominação (sinônimos), p.ex: curso d'água, rio, riacho, etc.

Algumas regras ligadas às propriedades (*fds*) que distinguem uma subclasse da sua superclasse também foram incorporadas ao PRONTO© como axiomas, todos dependentes do elemento básico dos módulos de edição e de cálculo do PRONTO© - o *consin*:

- ***Axioma 4:*** As *fds* de cada classe são representadas pelo conjunto {partes, funções e atributos};
- ***Axioma 5:*** As partes são elementos estruturais de uma classe;
- ***Axioma 6:*** As partes devem ser associadas a substantivos simples ou compostos (V. *Axioma 2*);

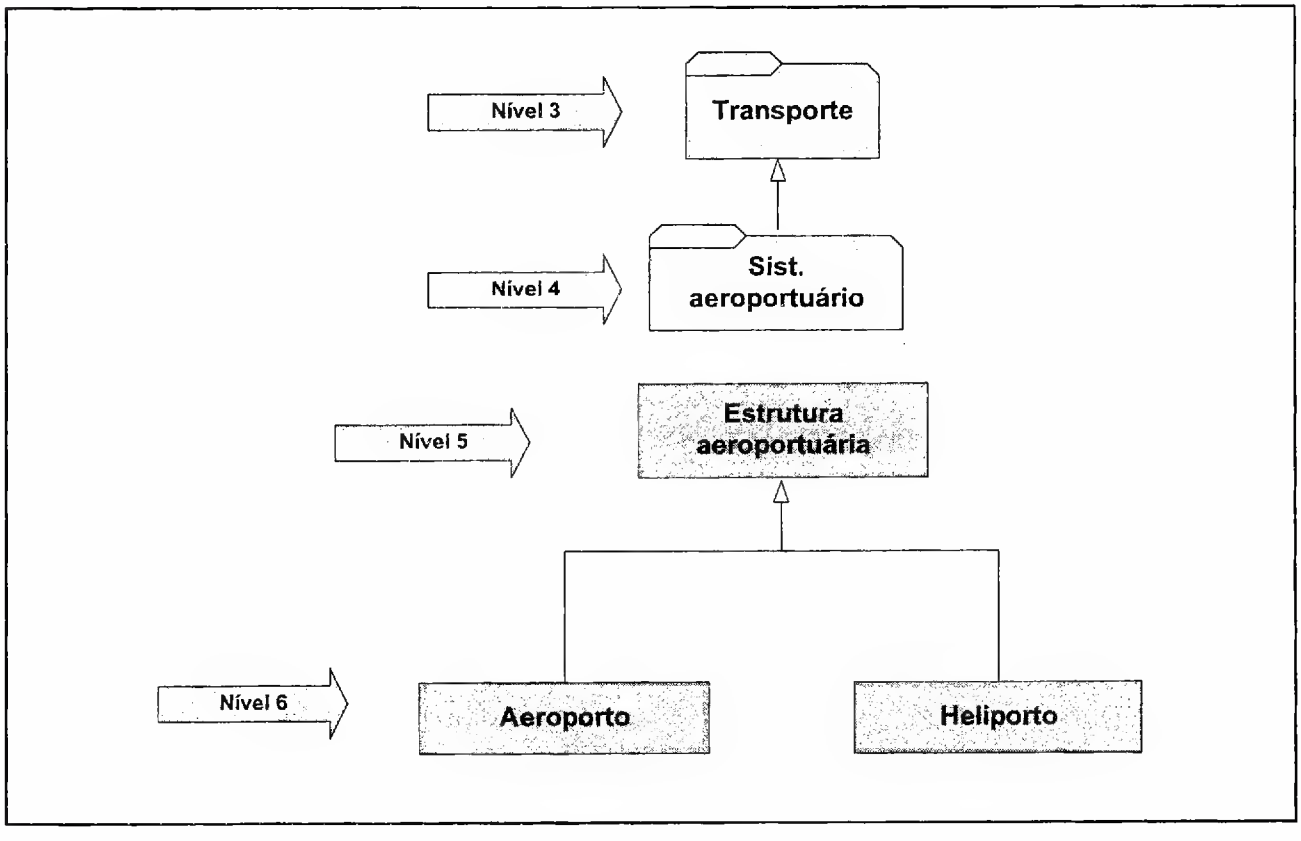

Figura 6.14 : Relações de generalização e agregação para a subcategoria sistema de transporte aeroportuário.

- ***Axioma 7:*** As funções devem ser associadas a verbos concretos;
- ***Axioma 8:*** Cada verbo deve manter uma correspondência biunívoca com a função que representa, ou se referindo a uma ação ou comportamento desempenhado por uma classe, ou se referindo a uma ação aplicada sobre esta classe;

- ***Axioma 9:*** Os atributos são propriedades adicionais de uma classe, não se enquadrando nem como funções e nem como partes;
- ***Axioma 10:*** Aos atributos se aplicam as mesmas regras de formação das partes (V. *Axiomas 2* e *6*);***Axioma 11:*** Os nomes das classes, suas relações de generalização, suas relações partitivas e as *fds* são representadas no PRONTO© por *consins*;
- ***Axioma 12:*** Os *consins* são comparados entre si por exaustiva operação de concatenação de cadeias de caracteres ou *strings* (V. *Axioma 2* para pontuação);
- ***Axioma 13:*** não são permitidos caracteres especiais como ´% & # $ / \ para atribuir nomes aos termos e também para desenvolver o texto das definições no interior do módulo de edição do PRONTO©.

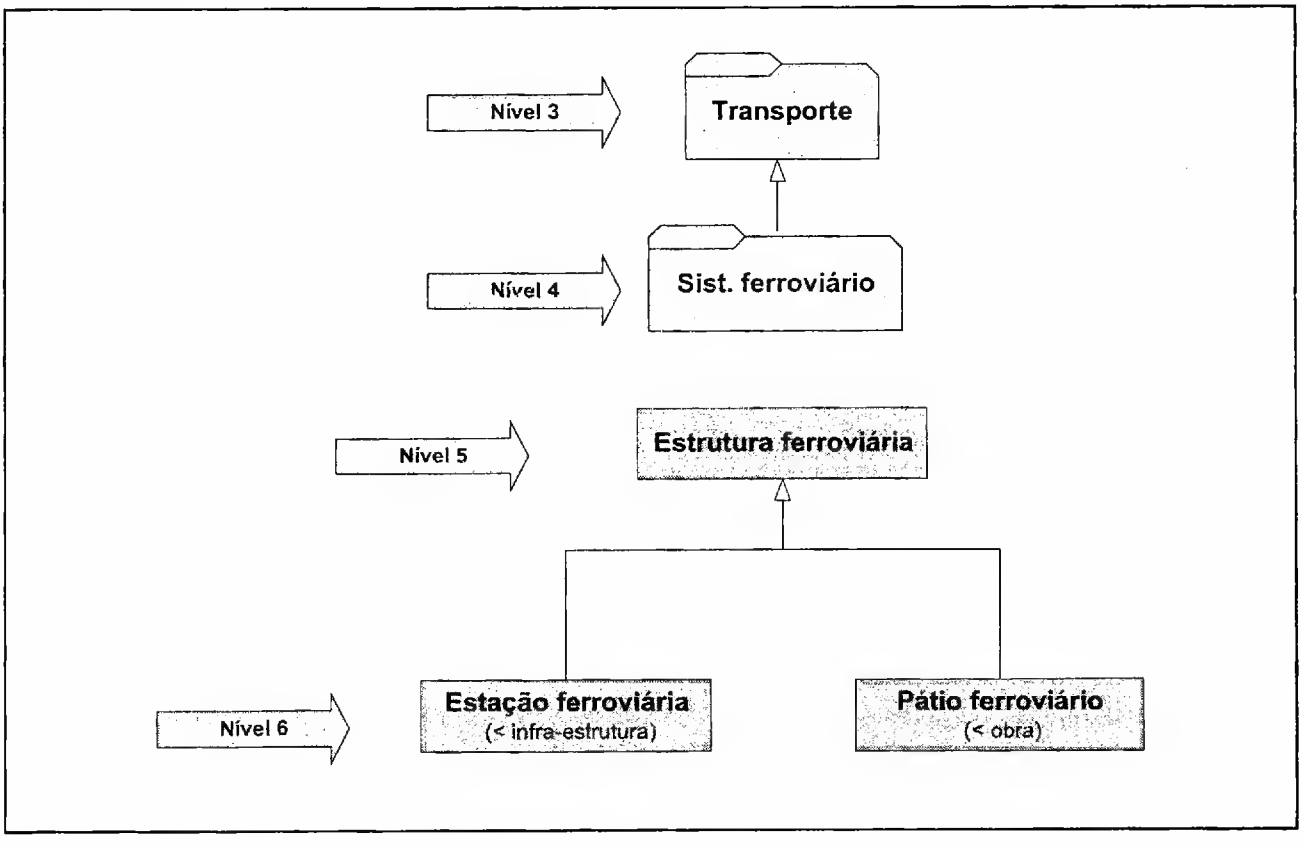

Figura 6.15: Relações de generalização e agregação para a subcategoria sistema de transporte ferroviário.

Pode haver certa confusão entre as partes (feições distintivas) e uma classe que seja parte de uma outra classe, numa relação partitiva ou meronímica. RODRÍGUEZ (2000) resolveu a dúvida ao enunciar o que o PRONTO© incorporou como o seu *Axioma 5*, que não faz parte orgânica da implementação do protótipo, mas é dependente de supervisão do usuário humano; p.ex: *telhado e piso* são *partes* (*fds*) de *edifício*, mas *escola* e *bloco residencial*

são *componentes* (partes ou merônimos) de *edifício*. As primeiras não definem uma classe. As segundas têm o poder de categorizar ou definir, visto que são outras classes.

Os sinônimos entram no processo de casamento ou concatenação das *fds*. Cada termo é tratado da mesma forma em relação ao seu sinônimo. Essa operação de concatenação foi relatada por RODRÍGUEZ (2002) e contribui para o que a autora denominou de *similaridade restrita*; porquanto, é útil para reduzir a polissemia em ontologias que apresentam definições muito extensas.

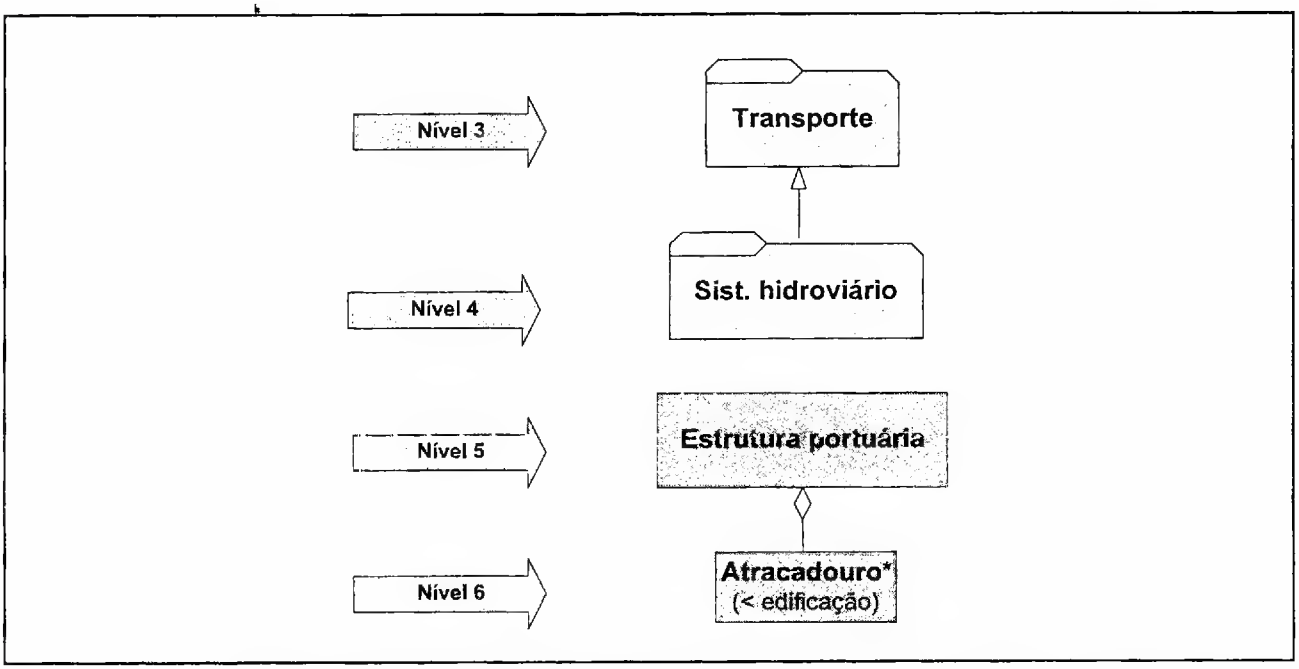

Figura 6.16: Relações de generalização e agregação para a subcategoria sistema de transporte hidroviário.

O nó-raiz (nível 1) da taxinomia foi atribuído à FOLHA FAXINAL, que abrange as oito categorias fundamentais do MC desta folha[268] (Figura 6.9), todas no nível 2 de abstração.

Neste ponto, vale fazer um apontamento: o MC da folha Faxinal e as TBCD©, quanto às grandes categorias temáticas, não coincidem quantitativa e qualitativamente. A diferença é pouca, mas enquanto o MC conta com oito grandes categorias, as TBCD© possuem nove grandes categorias. Na ordem alfabética, a atual versão das TBCD (2002) conta com: 1. *Altimetria* (RELEVO da antiga), 2. *Edificações*, 3. *Hidrografia*, 4. *Infra-estrutura*, 5. *Limites*, 6. *Localidades*, 7. *Pontos de Referência*, 8. *Sistema de Transporte* e 9. *Vegetação*.

[268] V. TBCD© no subitem 3.1.1 e no glossário, porque este MC foi baseado, em parte, na construção dessas tabelas.

Os comentários e avaliações que se seguirão sobre o MC da folha Faxinal, de fundamentação OO, servem como prática antecipada de Engenharia Ontológica, mercê do exercício de crítica às abstrações levadas a termo, especialmente no caso da categoria TOPONÍMIA, como se verá adiante. Portanto, daqui para frente, o foco do estudo será sobre o subconjunto do MC da folha Faxinal, ilustrada pela Figura 6.9 (oito grandes categorias).

Algumas categorias (EDIFICAÇÃO, LIMITE e TOPONÍMIA) foram hachuradas na Figura 6.9 e é necessário explicar o porquê.

Em relação às categorias EDIFICAÇÃO e LIMITE, as hachuras se devem simplesmente porque ambas não foram utilizadas nos experimentos de avaliação de SS.

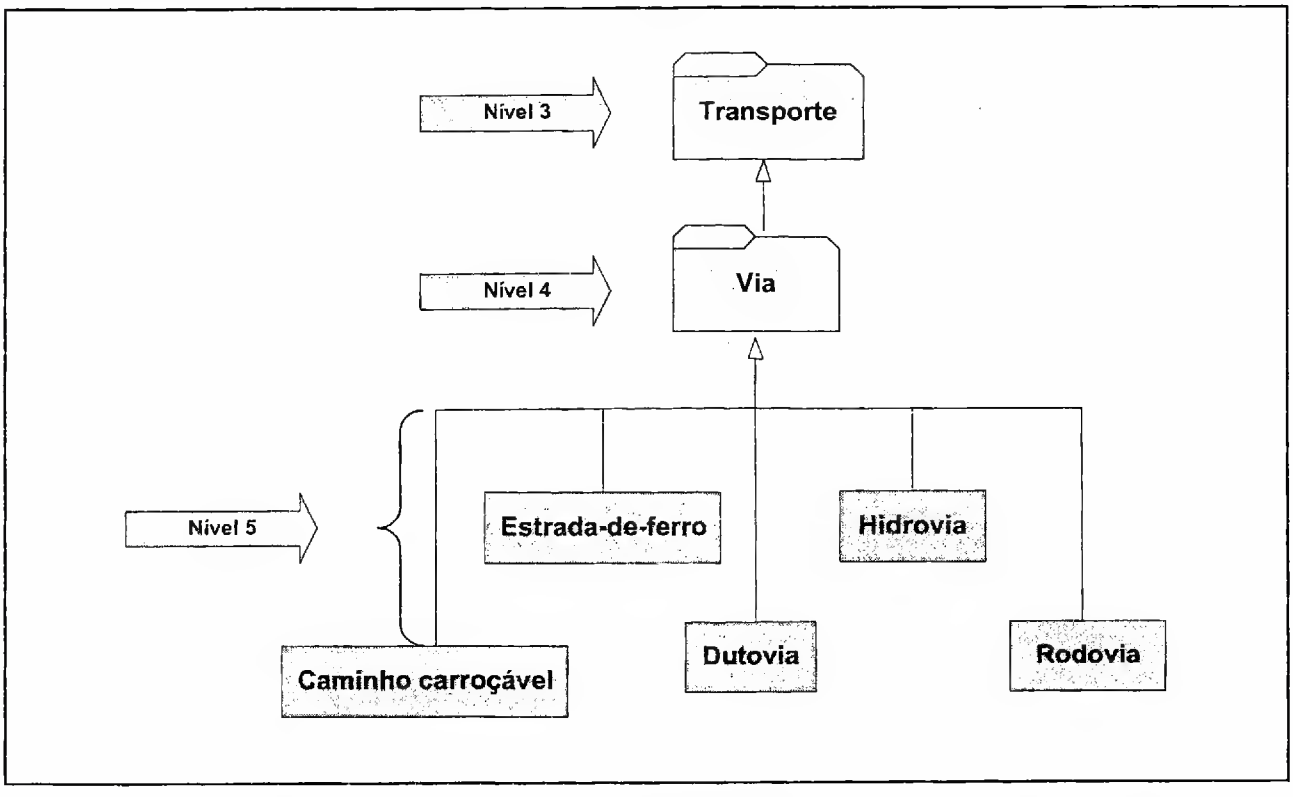

Figura 6.17 : Relações de generalização e agregação para a subcategoria via de transporte.

No caso de TOPONÍMIA, além da razão anterior, há mais o que explicar. É de consenso generalizado na comunidade cartográfica nacional[269] (BRASIL, 1998b) que a toponímia é um atributo de uma classe de entidades espaciais e, como tal, não pode constituir uma classe independente num modelo conceitual como o da folha Faxinal. Esse equívoco, no entanto, não foi cometido na versão de 2002 das TBCD©, mas é interessante comentá-lo como exercício de pesquisa.

---

[269] V. CEPAD/CONCAR no glossário.

Não só pela OO o equívoco de instituir TOPONÍMIA como uma categoria no MC da folha Faxinal se manifesta, mas até mesmo pela Terminologia, já que os topônimos são elementos terminológicos semelhantes a outros elementos gráficos que não possuem características, i.e., não podem participar de um nó de uma rede semântica (sistema de conceitos), como a que se pretende produzir para os termos geográficos da folha Faxinal. Esses topônimos (antropônimos) são utilizados como elementos construtores (realizações terminológicas) de definições de classes de entidades espaciais e, dessa forma, não podem contribuir com características de categorização do modelo conceitual geográfico.

Se o domínio de estudos fosse restrito aos elementos estruturais gráficos (*grafismo*) da folha Faxinal, a categoria TOPONÍMIA seria representativa do esforço de abstração para modelar a essência do fenômeno gráfico em questão. No entanto, não é esse o caso, pelo menos nessa fase da gênese de um SIG, em que o foco da análise deve se voltar para a visão de um mundo que vai muito além do mero *grafismo* (simbolização final dos fenômenos geográficos). A "visão" desse MC deve orientar-se para as entidades espaciais genuínas que "povoam" o mundo real e para a descoberta das ações e reações a que estão sujeitas.

Algumas notas quanto à denominação das categorias dos diagramas:

- Para uniformizar a nomenclatura, todas as categorias do MC original foram renomeadas no singular;
- LIMITE, em vez de REFERENCIAL (do original), é o termo com maior poder de denotação para marcas em geral: marcos, pilares e linhas divisórias de áreas. "Referencial" é um sistema de eixos coordenados que permite localizar a posição de uma coisa no espaço (LELLO, 1984).

Na versão de 2002 das TBCD©, uma instanciação da grande categoria LIMITE foi "promovida" na escala hierárquica. Trata-se da classe *Pontos de Referência*, destinada à materialização de elementos pontuais do terreno, de forma dinâmica ou estática. Com isso, *Limites*, na atual versão, contempla a representação de elementos lineares de divisa ou de demarcação.

Além do que já foi descrito para a categoria LIMITE e TOPONÍMIA, é necessário juntar ao texto descrições sucintas das outras seis categorias do MC da folha Faxinal (V. Figura 6.9), o que vai ao encontro das expectativas do público que não é afeito às particularidades da Engenharia Cartográfica. A fonte de consulta para as descrições em tela ficou restrita às TBCD© e ao manual técnico T 34-700 (BRASIL, 1998a).

A categoria EDIFICAÇÃO engloba classes de construções humanas não classificadas na categoria INFRA-ESTRUTURA. A característica básica dessa grande categoria está centrada no uso e finalidade social, cultural e particular dessas construções.

A categoria HIDROGRAFIA (Figura 6.10) engloba o conjunto das águas correntes ou estáveis, intermitentes ou regulares de uma região, além das entidades espaciais naturais ou artificiais, expostas ou submersas, que estejam contidas nesse ambiente.

A grande categoria INFRA-ESTRUTURA (Figura 6.11) abrange quatro subcategorias de nível 3 e mais seis subcategorias de nível 4, derivadas das de nível 3 (Figuras 6.12 a 6.19). É a categoria que engloba a base material e econômica nas áreas de indústria de base, energia, mineração, extrativismo mineral, comunicação, saúde, educação, saneamento, irrigação, lazer ou áreas em que estejam sendo desenvolvidas atividades relevantes, tanto pela iniciativa governamental como pela privada, para o desenvolvimento de uma dada região, sempre que essas atividades tenham por meta atender às necessidades sociais.

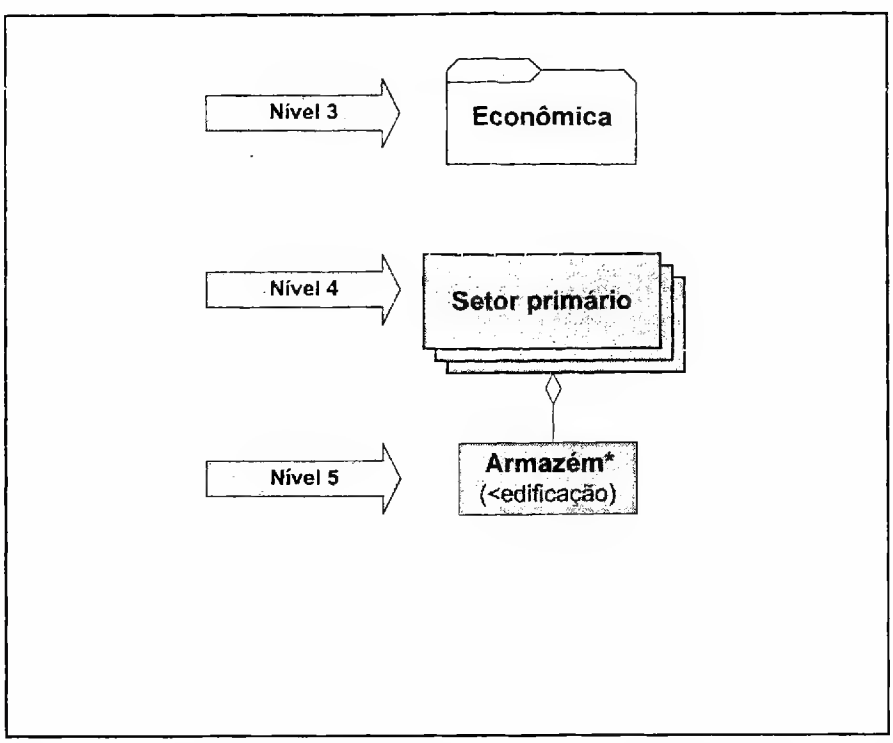

Figura 6.18 : Relações de generalização e agregação para a subcategoria infra-estrutura econômica.

A categoria LOCALIDADE (Figura 6.20) engloba os elementos espaciais que definem os tipos e áreas de ocupação humana, classificados segundo a legislação em vigor.

A categoria RELEVO (Figura 6.21) engloba os aspectos morfológicos do terreno.

A categoria VEGETAÇÃO (Figura 6.22) engloba as espécies vegetais naturais ou cultivadas, classificadas quanto ao seu porte ou ciclo produtivo, respectivamente.

As definições formuladas para os questionários (Apêndice A) aplicados aos indivíduos da DSG e do CCAuEx não correspondem às definições carregadas na ontologia *ad-hoc*.

As definições dos questionários não seguiram as métricas da *estrutura formal da definição predicativa*. Esse relaxamento foi conveniente porque poupou tempo na sua formulação e porque se utiliza de exemplos, o que torna o texto mais assimilável ao respondente.

O mesmo relaxamento é inconcebível no segundo caso de definição. Qualquer falha de controle terminológico nessa fase é capaz de produzir resultados imprevisíveis na fase de cálculo da SS. Virtualmente, os erros se potencializam para as fases subseqüentes de avaliação da SS e o modelo ficaria comprometido em termos de consistência.

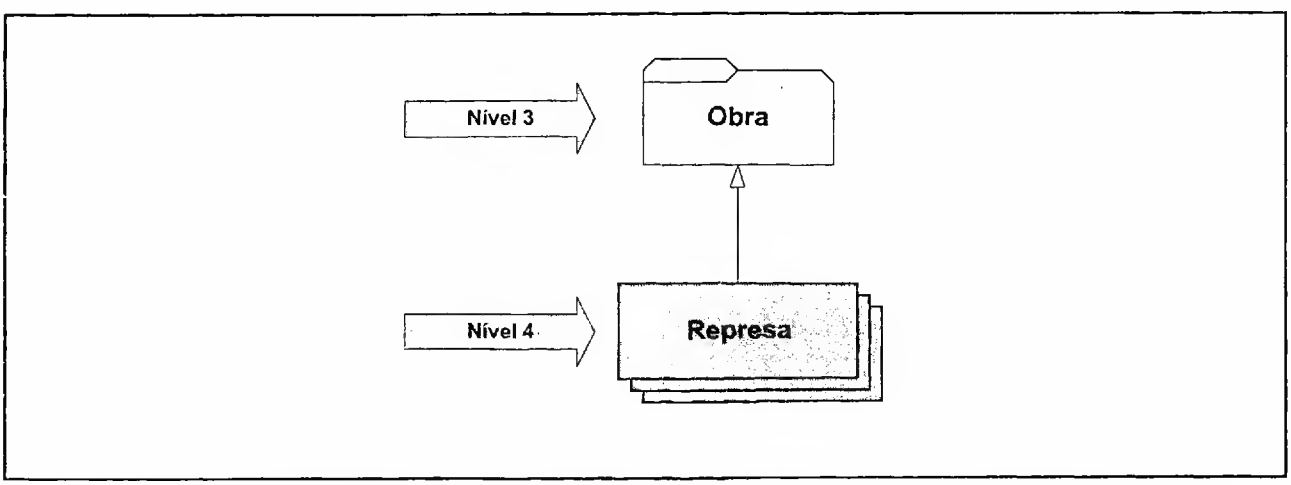

Figura 6.19: Relações de generalização e agregação para a subcategoria obra.

Segundo a *estrutura formal da definição predicativa,* definição conotativa ou ainda definição *intensional*, o princípio básico é definir um termo com base no *gênero próximo e nas diferenças específicas*; p.ex: não é admissível definir "mesa" como "objeto de uso doméstico" ou como "um móvel de sala de jantar". No primeiro caso, o gênero é demasiadamente amplo, porque inclui um grande número de outros objetos que nada têm a ver com mesa. No segundo caso, o gênero é demasiadamente restrito, porque exclui outras espécies de mesa: mesa de cozinha, mesa de escritório, mesa de centro, etc. (GARCIA, 1976, p. 309).

Noutras palavras, o princípio da definição conotativa deve pautar-se por ser suficientemente amplo para compreender a espécie definida e suficientemente restrito para que as características individualizadas possam ser percebidas sem dificuldade e confusão com ou-

tras espécies. Dessa forma, uma definição mais adequada para "mesa", no exemplo anterior, poderia ser: "É um móvel composto de um tampo e de um meio para sustentá-lo".

Outras métricas assessórias da definição conotativa [GARCIA (1976, p. 309) e RUDIO (1978)] foram observadas quando foram montaram as tabelas (notação BNF) para carregar a ontologia no PRONTO©, entre elas:

- Estrutura gramatical rígida (sujeito e gênero da mesma classe gramatical);
- Formular uma definição científica implica usar conceitos adequados;
- Conceito adequado é aquele que é claro e distinto;
- Conceito claro é aquele que permite reconhecer a coisa a que se refere; é aquele que não é obscuro e que se expressa numa linguagem simples;
- Conceito distinto é aquele que permite distinguir as propriedades da coisa a que se refere; é aquele que não é confuso, i.e., que é preciso;
- Sentenças obrigatoriamente na forma afirmativa e direta (sujeito – verbo - complemento);
- A definição científica deve ser concisa (caber num só período).

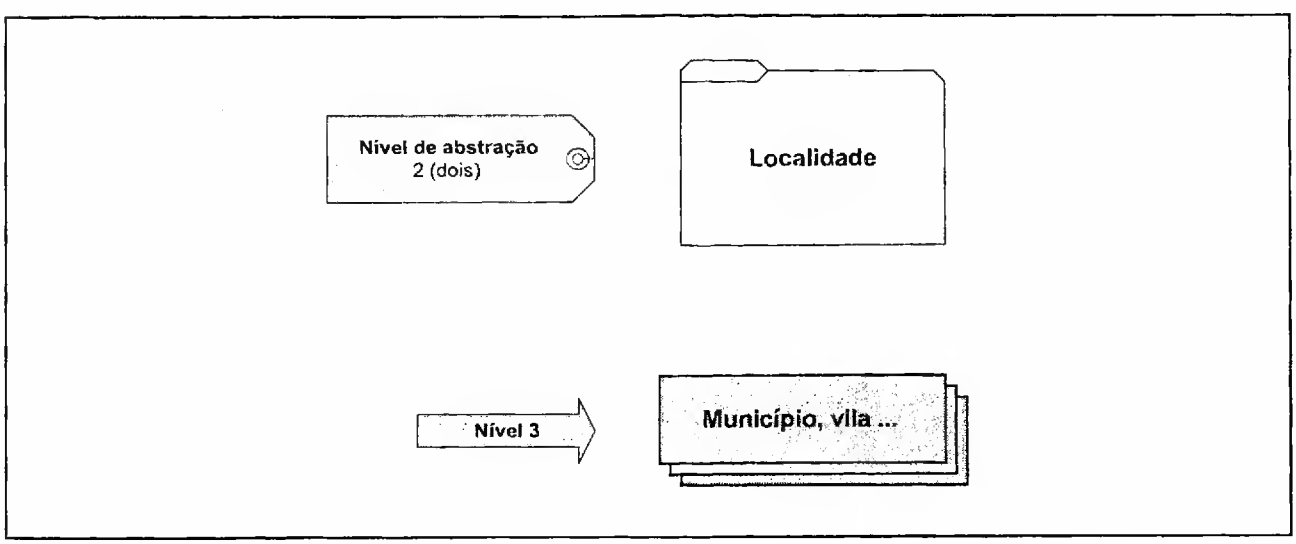

Figura 6.20: Relações de generalização e agregação para a categoria localidade.

As TBCD© e o manual T 34-700 (BRASIL, 1998a) foram as fontes lexicográficas básicas para as definições conotativas que compuseram a ontologia *ad-hoc*, na falta de documentação do MC da folha Faxinal. Todavia, nem sempre essas duas fontes cobriram todas as necessidades em conceitos para essa etapa da pesquisa, o que obrigou a se lançar mão de outras fontes, conforme já relatado.

A nota anterior tem o intuito de alertar para o fato de que há algumas diferenças entre os modelos conceituais das fontes; p.ex., nas TBCD©, "rio" é sinônimo de "curso d´água" e a ontologia *ad-hoc* foi estruturada nesse sentido. Na lexicografia *on-line Wordnet*™, no entanto, a que foi utilizada por RODRÍGUEZ (2000) para montar a sua ontologia, "rio" é uma instância de "curso d´água". Em que pese essas diferenças de organização hierárquica do conhecimento, tanto o resultado obtido pela autora na sua avaliação de SS pelo MSS como o obtido pelo PRONTO©, como se verá a seguir, foram satisfatórios, o que parece certificar o conceito de ontologia como a especificação formal (explicitação) de uma conceituação compartilhada, de acordo com vários autores como N. Guarino e T. Gruber [*apud* LEÃO(2003)], USCHOLD (1996) e KIRYAKOV (1998).

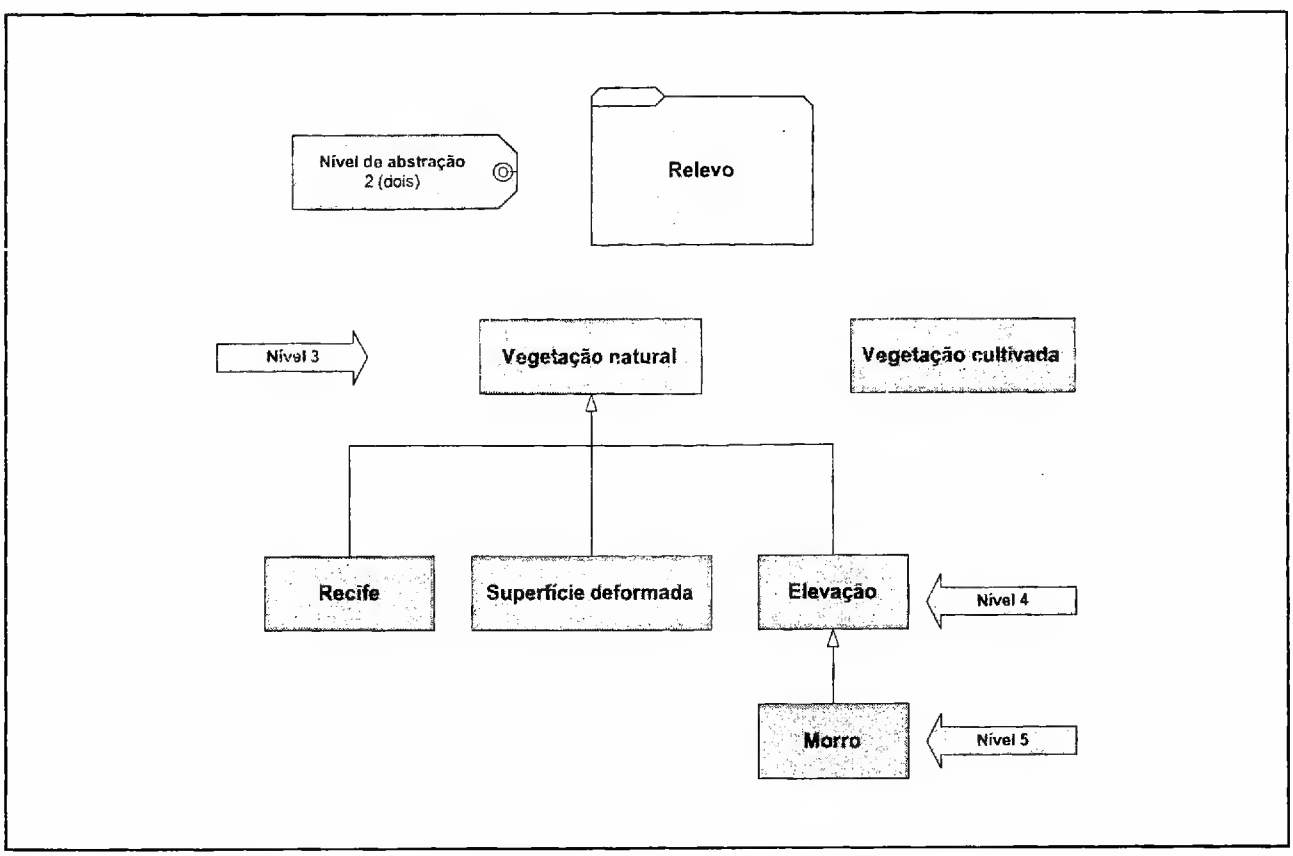

Figura 6.21: Relações de generalização e agregação para a categoria relevo.

Os resultados satisfatórios comentados anteriormente são muito dependentes do modelo conceitual adotado e não é difícil perceber que a ontologia de RODRÍGUEZ (2000), baseada em lexicografias de utilidade geral, como a *Wordnet*™, complementada com termos espaciais de catálogos como o SDTS™, teve a sua plausibilidade confirmada pela visão do

mundo (*conceitualização*[270], segundo os autores do parágrafo anterior) que indivíduos de um curso universitário de Letras deixaram registrada no questionário que ela aplicou.

Já no caso desta pesquisa, a visão ou representação mental do mundo definido no questionário foi a de um grupo de indivíduos profissionais da área *geocientífica* e a ontologia que foi carregada no PRONTO© não foi orlada de fontes de utilidade geral, como no caso anterior, mas sim de um modelo conceitual concebido por profissionais dessa mesma área (para evitar tendenciosidade nas respostas, nessa fase da pesquisa, não se admitiu respondente que tivesse participado da elaboração do MC da folha Faxinal).

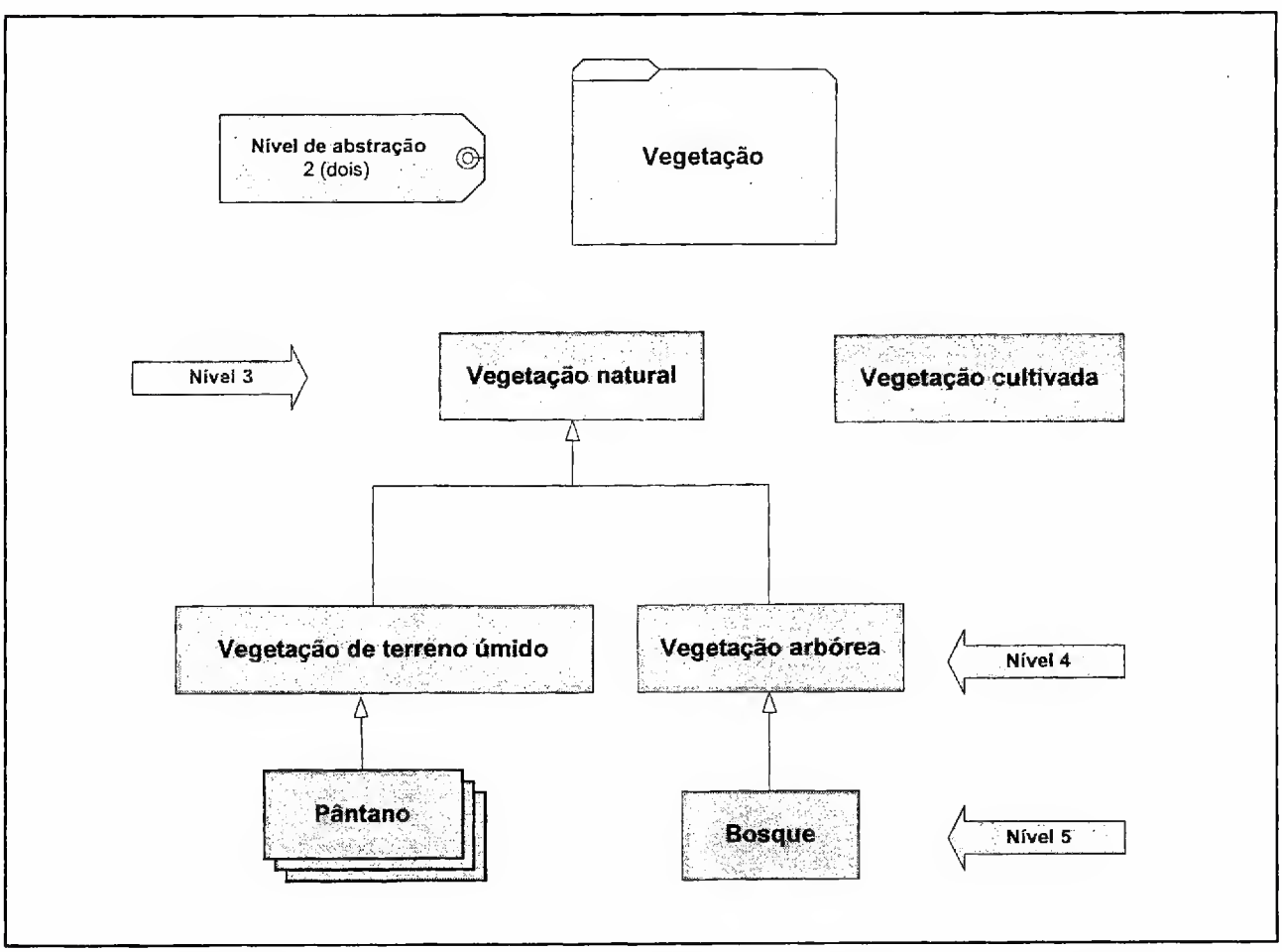

Figura 6.22: Relações de generalização e agregação para a categoria vegetação.

Essas correspondências entre as naturezas geral e específica das ontologias com os espaços de estimulação sensorial dos respondentes, talvez tenha garantido um resultado

[270] V. conceito de *esquema* na p. 146.

em média muito bom para a avaliação de SS entre as classes de entidades espaciais escolhidas para os dois casos de avaliação (PROFAX e PRONTO©), embora discrepâncias tenham ocorrido, como se observará no caso da classe PÂNTANO (subitem 6.4.2).

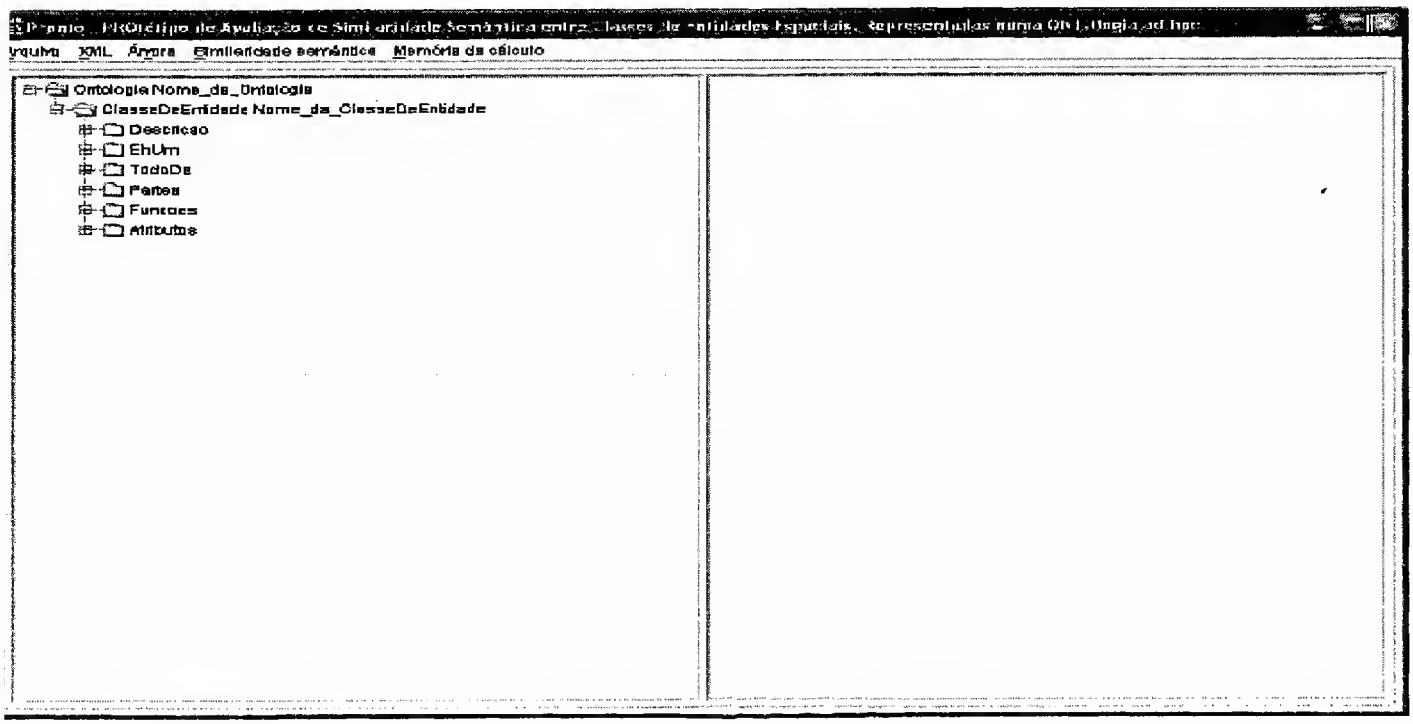

Figura 6.23: Tela de abertura do PRONTO.

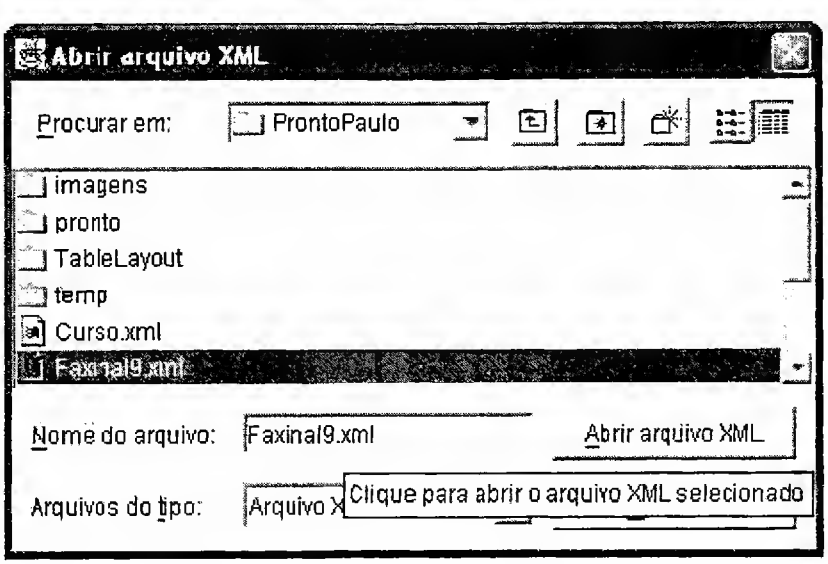

Figura 6.24: Módulo de abertura do arquivo XML que contém a taxinomia (árvore *n-ária*).

As Figuras 6.23 a 6.29, a seguir, ilustram uma seqüência de operações no PRONTO©, particularmente no que se refere à carga da ontologia *ad-hoc* e aos cálculos de SS para algumas classes de entidades espaciais que foram mostradas nos diagramas anteriores.

Neste ponto, torna-se produtivo encerrar a lista de axiomas da ontologia *ad-hoc* para a folha Faxinal. Além dos treze axiomas anteriores, a lista[271] ainda incorpora os seguintes:

- ***Axioma 14:*** A abstração lógica do PRONTO© chama-se "elemento";
- ***Axioma 15:*** Um elemento do tipo "ontologia" deve ser um conjunto não-vazio, i.e., conter pelo menos um elemento chamado "classe de entidades";

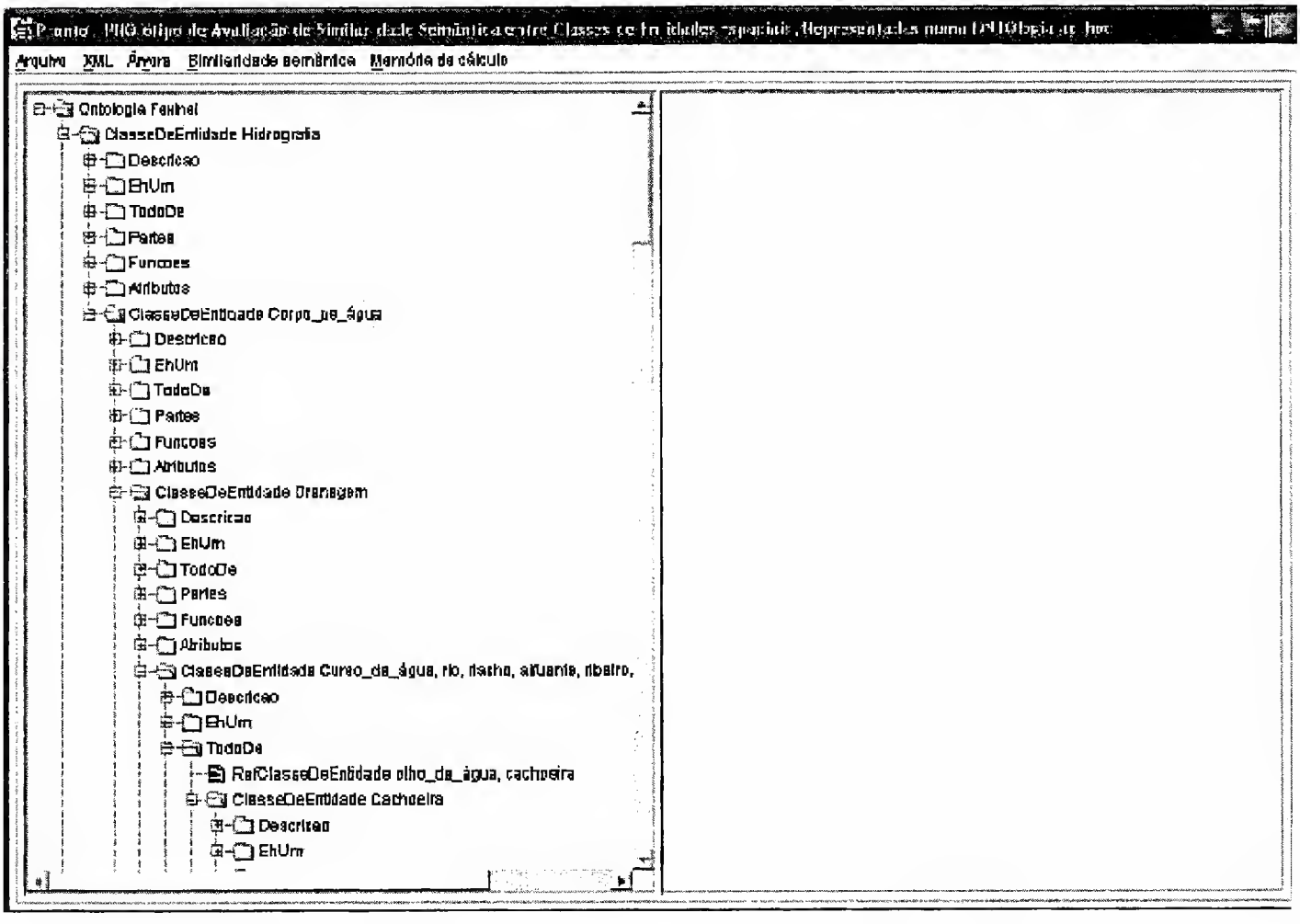

Figura 6.25: Interface gráfica de apresentação da árvore *n-ária*.

- ***Axioma 16:*** Cada classe de entidades deverá possuir:
  - Um elemento do tipo 'nome', o qual deverá possuir um elemento do tipo *'consin'*;

[271] A lista completa constitui o conteúdo do arquivo de texto *Ontologia.dtd*, que vai no disquete preso à capa final.

- Um elemento do tipo 'descrição', o qual deverá conter apenas informação na forma de caracteres PCDATA (*Parsed Character Data*), que poderá ser nula;
- Um elemento do tipo 'é-um', o qual deverá possuir apenas um elemento do tipo 'referência_classe_entidades';
- Um elemento do tipo 'todo-de', o qual poderá possuir nenhum (ou vários) elemento(s) do tipo 'referência_classe_entidades';

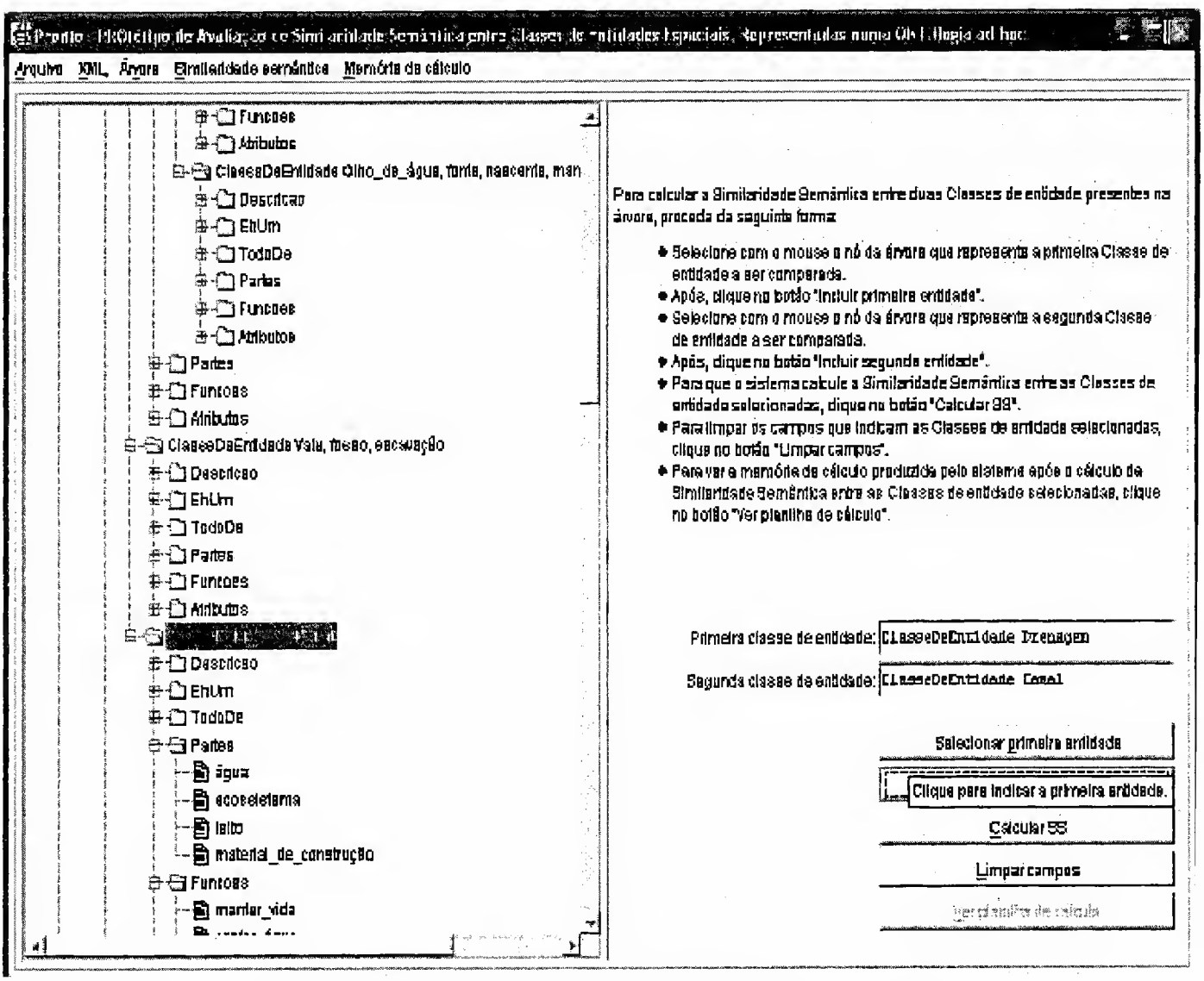

Figura 6.26: Seleção das classes DRENAGEM e CANAL para a avaliação de SS.

- Um elemento do tipo 'partes', o qual poderá possuir um ou vários elementos do tipo *'consin'*;
- Um elemento do tipo 'funções', o qual deverá possuir um ou vários elementos do tipo *'consin'*;

- Um elemento do tipo 'atributos', o qual poderá possuir um ou vários elementos do tipo *'consin'*.

- ***Axioma 17:*** O elemento 'nome' deverá possuir um elemento do tipo *'consin'*, ou seja, o elemento não poderá conter outros elementos, 'texto', ou referências a outras entidades ('referência_classe_entidades');
- ***Axioma 18:*** O elemento 'descrição' deverá conter apenas informação na forma de caracteres (PCDATA), ou seja, o elemento não poderá conter outros elementos ou referências a outras entidades;
- ***Axioma 19:*** O elemento 'é-um' deverá possuir apenas uma referência a um elemento do tipo 'referência_classe_entidades', que é também um elemento;

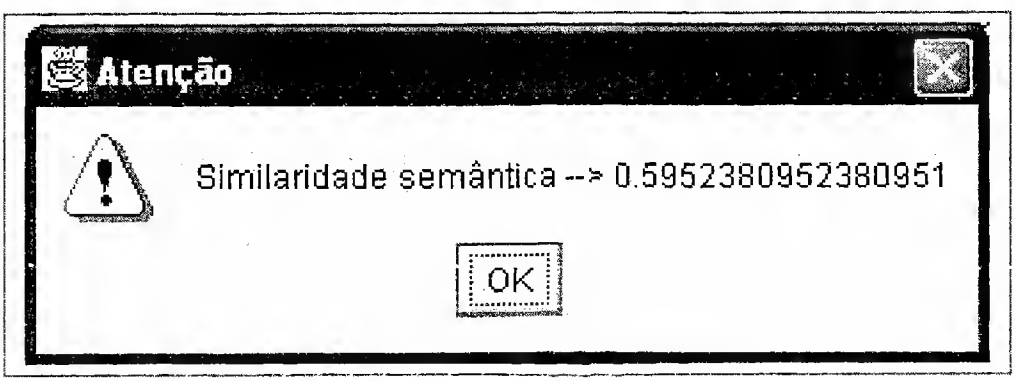

Figura 6.27: Cálculo da SS entre a classe DRENAGEM e a classe CANAL.

- ***Axioma 20:*** O elemento 'é-um' não poderá conter referências a outros tipos de elementos, 'texto' ou referências a outras entidades;
- ***Axioma 21:*** O elemento 'todo-de' deverá possuir apenas uma referência a um elemento do tipo 'referência_classe_entidades', que é também um elemento;
- ***Axioma 22:*** O elemento 'todo-de' não poderá conter referências a outros tipos de elementos, 'texto' ou referências a outras entidades;
- ***Axioma 23:*** O elemento 'partes' poderá possuir zero (nenhuma) ou várias referências a um elemento do tipo *'consin'*, que é também um elemento;
- ***Axioma 24:*** O elemento 'partes' não poderá conter referências a outros tipos de elemento, 'texto' ou referências a outras entidades;
- ***Axioma 25:*** O elemento 'funções' poderá possuir uma ou várias referências a um elemento do tipo *'consin'*, que é também um elemento;
- ***Axioma 26:*** O elemento 'funções' não poderá conter referências a outros tipos de elemento, 'texto' ou referências a outras entidades;

- ***Axioma 27:*** O elemento 'atributos' poderá possuir nenhuma ou várias referências a um elemento do tipo *'consin'*, que é também um elemento;
- ***Axioma 28:*** O elemento 'atributos' não poderá conter referências a outros tipos de elemento, 'texto' ou referências a outras entidades;

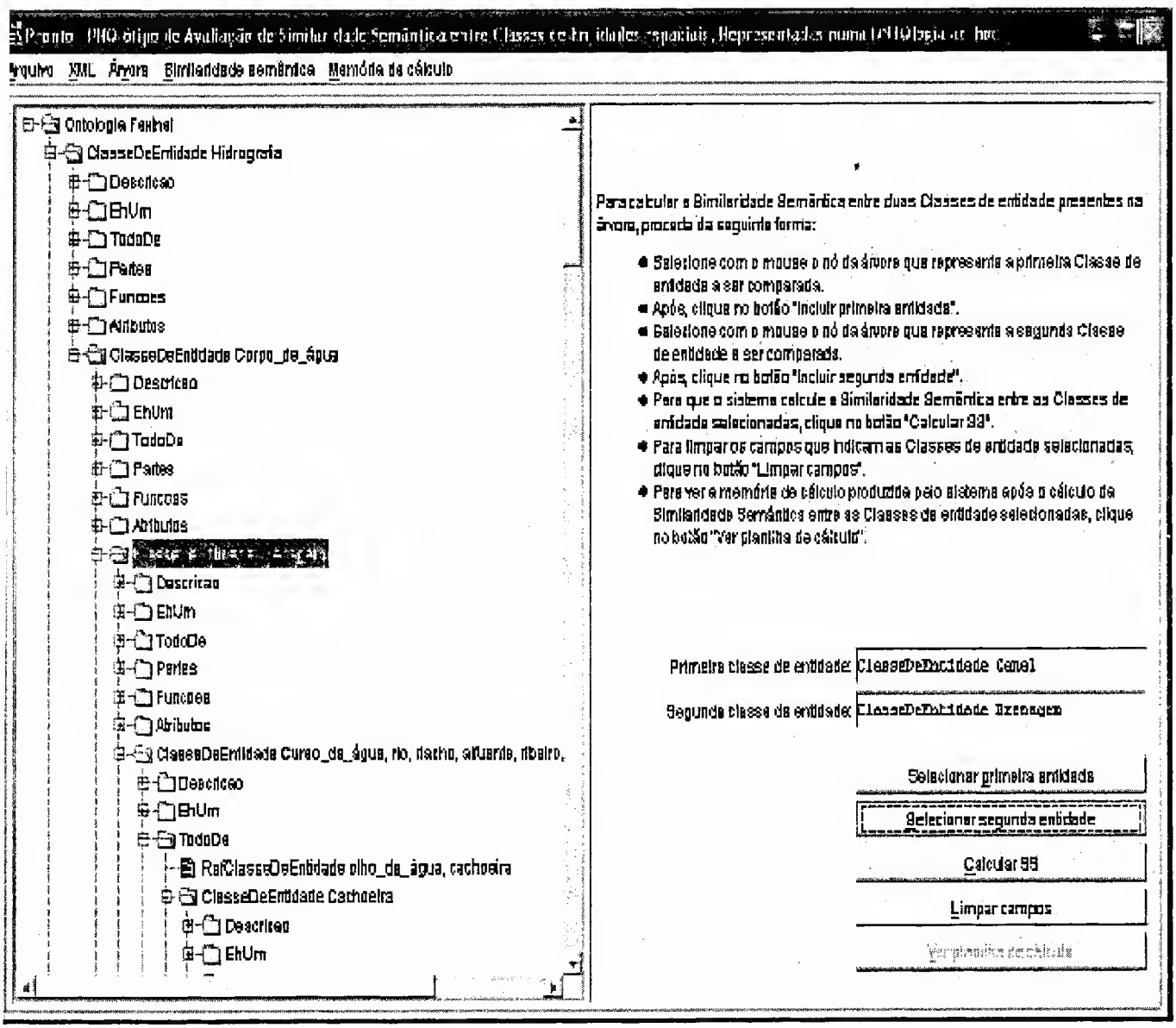

Figura 6.28: Seleção das classes CANAL e DRENAGEM para a avaliação de SS.

- ***Axioma 29:*** O elemento *'consin'* deverá conter apenas informação na forma de caracteres, ou seja, o elemento não poderá conter outros elementos, 'texto', ou referências a outras entidades;
- ***Axioma 30:*** O elemento 'referência_classe_entidades' deverá possuir uma ou várias referências a um elemento do tipo 'nome', que é também um elemento;

- ***Axioma 31:*** O elemento 'referência_classe_entidades' não poderá conter referências a outros tipos de elemento, 'texto ou outras referências a entidades.

As Figuras 6.23 e 6.24 constituem a testificação implementada do Axioma 15. Este axioma coincide em teor com um princípio fundamental da OO e, por extensão, da Lógica, em que não há significado num sistema de conceitos para uma classe de entidades vazia. A classe deve possuir, no mínimo, uma função.

A Figura 6.23 mostra a classe fundamental para o usuário que deseje construir a sua ontologia. É a matriz de uma classe de entidades (espaciais ou não) que surge automaticamente, assim que o PRONTO© é aberto.

A Figura 6.24 mostra o preparo para a carga (Figura 6.25) de uma ontologia previamente construída para fornecer os termos (referentes de classes de entidades espaciais) que serão transferidos para o módulo de cálculo de SS.

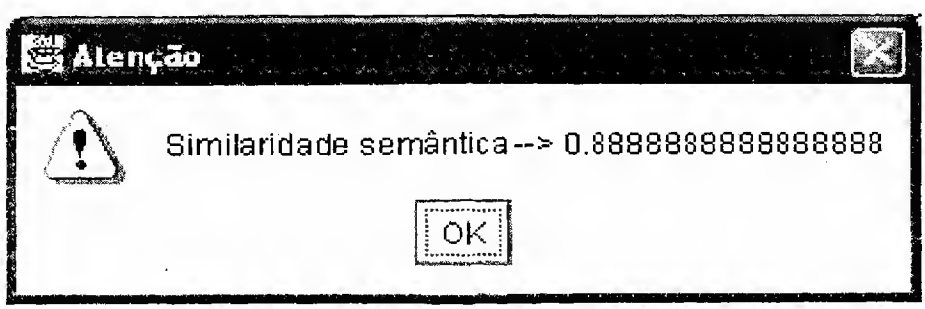

Figura 6.29: Cálculo da SS entre a classe CANAL e a classe DRENAGEM.

A Figura 6.26 mostra o preparo para o cálculo da SS entre uma classe de um nível mais alto de generalização (sujeito ou variante: DRENAGEM) e uma de um nível mais baixo (predicado ou protótipo: CANAL).

A Figura 6.27 mostra o resultado da SS entre DRENAGEM e CANAL (0,595).

A Figura 6.28 mostra o preparo para o cálculo da SS entre uma classe de um nível mais baixo de generalização (sujeito ou variante:CANAL) e uma de um nível mais alto (predicado ou protótipo: DRENAGEM).

O resultado da SS entre canal e DRENAGEM é ilustrado na Figura 6.29 (0,888).

Esses resultados confirmam o que já se presumia no exemplo de treinamento, em que foi efetuada a carga no protótipo de uma ontologia bem simples e fraca em termos conceituais. Aqui, dispondo-se de uma ontologia mais complexa e formalizada de maneira a não se afastar muito do MR, demonstrou-se que este modelo conceitual, implementado pelo

PRONTO©, simula coerentemente o efeito da assimetria e que oferece comodidade para a construção de ontologias. É preciso, agora, conhecer o valor estatístico dessas medidas realizadas pelo PRONTO©.

6.3.1.3. A linguagem de implementação dos protótipos

A LTP *Java*™ foi a linguagem de programação adotada para a implementação.

O subitem 3.2.2.2.3 cobre as considerações que levaram à escolha dessa linguagem técnica de programação para ambos os protótipos.

O CD que acompanha este volume (Apêndice B) contém, além do código-fonte do PRONTO© e o arquivo da ontologia *ad-hoc* (Faxinal9a.xml), os diagramas de classes, o de caso-de-uso, entre outros documentos (arquivo leia_me.doc, tabelas do SPSS™, etc.)

6.3.2. Metodologia estatística

"Use a estatística do mesmo modo que um ébrio; para ele o poste vale mais pelo apoio do que propriamente pela iluminação." [Andrew Lang (*apud* WONNACOTT (1980, p. 3)]

Foram empregados os métodos não-paramétricos para corroborar as duas hipóteses estatísticas formuladas.

Segundo DANIEL (1978), WONNACOTT (1980, p. 171), LEVIN (1987) e SIEGEL (1988), os estimadores de tendência central (média e mediana) e de dispersão (variância) são adequados para cálculos estatísticos em que as populações seguem distribuições que se aproximem da forma normal ou de outra qualquer conhecida.

No entanto, quando a forma da distribuição da população da qual se extraem as amostras é desconhecida, já não é mais evidente que estimador é o mais adequado. Logo, é veementemente recomendável o uso de um estimador que seja razoavelmente eficiente para qualquer forma de população. Tal estimador, que não depende da hipótese de a distribuição ser normal, é chamado de *robusto*, *livre de distribuição* ou ***não-paramétrico***.

Os métodos estatísticos não-paramétricos são aqueles em que o pesquisador não pode, honestamente, admitir normalidade de distribuição para os seus dados, i.e., os dados podem estar assimetricamente distribuídos, especialmente quando se trata de dados colhidos em níveis nominais ou ordinais de mensuração, muito comuns no âmbito das ciências sociais e nas ciências do homem (Psicologia, p.ex.), conforme explica LEVIN (1987, p.194).

Foram dois os testes não-paramétricos utilizados neste experimento:

- Teste de Concordância de *Kendall* (estimador: coeficiente *W*)
- Teste de Correlação de *Spearman* (estimador: coeficiente R_s)

As fórmulas de cálculo dos coeficientes *W* e R_s são as expressas pelas Equações 6.4 e 6.5, respectivamente.

No caso desta pesquisa, essas equações foram resolvidas pelo aplicativo estatístico SPSS™, que se reveste numa poderosa ferramenta de cálculo e de auxílio no planejamento de experimentos que necessitam de apoio da Estatística, cobrindo as etapas que vão desde a coleta até a interpretação dos resultados, passando pela transformação e análise de dados.

$$W = \frac{12\sum R_i^2 - 3N^2 m(m+1)^2}{N^2 m(m^2-1)} \qquad \text{Eq. 6.4}$$

Na Equação 6.4, ***N*** representa o número de casos (respostas válidas dos indivíduos, p.ex.), ***m*** representa o número de variáveis (classes de entidades julgadas, p.ex.) e ΣR_i^2 representa o somatório dos postos para cada ***m*** variáveis ordenadas.

$$R_s = 1 - \frac{\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \qquad \text{Eq. 6.5}$$

Na Equação 6.5, Σd_i^2 representa o somatório das diferenças observadas entre os postos de duas variáveis ***x*** e ***y*** e ***N*** constitui o número de pares ***(x, y)*** dos dados.

Vários autores estatísticos comentam os casos de uso de um ou de outro estimador (*W* de Kendall e R_s de Spearman) para situações experimentais específicas. Entre esses autores, DANIEL (1978) e SIEGEL (1988) são claros quanto ao fato de que os dois coeficientes, mesmo aplicados aos mesmos conjuntos de dados, não podem ser numérica e diretamente comparados entre si, apesar de constituírem estimativas de como duas ou mais variáveis alteram seus conteúdos conjuntamente[272] (correlação). Eles também concordam em alguns pontos, a seguir resumidos:

- O coeficiente R_s é usado para calcular a correlação entre amostras de dados extraídos de populações *bivariadas*, em que os valores assumidos pelas duas variáveis (x, y) são ordenadas por um critério comum, p.ex: quando são categorizadas por postos;

[272] A tendência do fenômeno.

- Os dados correspondentes às duas variáveis na correlação de *Spearman* devem ser de um tipo (mesmo não-numérico) que possa ser categorizado por níveis médios de quantificação, p.ex., o ordinal (por postos);
- R_s é um valor do domínio dos números reais, que varia no intervalo [-1, +1], em que -1 representa uma correlação perfeitamente negativa e +1, uma correlação perfeitamente positiva (0 indica que nada se pode afirmar);
- O coeficiente *W* é um número real que varia no intervalo [0, 1];
- O coeficiente *W* é indicado para amostras extraídas de populações multivariadas, i.e., com ***N*** casos ou unidades de observação, capazes de serem ordenadas por postos;
- O coeficiente *W* é o ideal para estudos que envolvam julgamentos e classificação de padrões de comportamento entre vários indivíduos, assim como em aplicações que necessitem categorizar variáveis (PCA e AF – V. subitem 1.2.3, notas de rodapé 26 e 27);
- Apesar de quantitativamente distintos, há fórmulas[273] que relacionam o coeficiente *W* e a média dos R_s, tomadas de um conjunto extenso de dados (SIEGEL, 1988, p.262);
- Ambos os coeficientes (*W* e R_s) produzem informações de correlação dos mesmos conjuntos de dados e, dentro de suas características, possuem a mesma capacidade de detectar o grau de associação a que estão submetidas as observações de uma dada população.

Ambos os estimadores não-paramétricos necessitaram ser testados quanto ao seu poder de discernimento estatístico e o tamanho das amostras foi determinante nesse aspecto, para que se pudesse aproximar o cálculo dos métodos tradicionais empregados em distribuições do tipo normal ou quiquadrado.

Segundo SIEGEL (1988, p. 244), é admissível estimar R_s por intermédio da variável normal padronizada **z**, se o tamanho da amostra exceder a cinqüenta casos[274]. A Equação 6.6 formaliza esta condição para que se possa estabelecer intervalos de confiança para R_s, i.e., testar estatisticamente esse estimador não-paramétrico.

$$z = R_s \sqrt{N-1} \qquad \text{Eq. 6.6}$$

Segundo RODRÍGUEZ (2000, p. 92), SIEGEL (1988, p. 269), LEVIN (1987, p. 220) e DANIEL (1978, p. 329), o teste quiquadrado serve para determinar o intervalo de confiança

---

[273] Não serão exploradas neste estudo.
[274] Para DANIEL (1978, p.304) o tamanho da amostra deve exceder 30 unidades de observação.

para o coeficiente *W*, conforme a Equação 6.7, em que "m" representa os graus de liberdade para o teste X^2 e "N" o número de casos válidos.

$$\chi^2 = N(m-1)W \qquad \text{Eq. 6.7}$$

Os mesmos autores acima, ainda estabeleceram condições para a aplicação do teste quiquadrado como forma de avaliar o nível de significância de *W*. Ei-las:

- Deve ser usado para levar o pesquisador a decidir se, numa amostra aleatória, duas ou mais variáveis guardam entre si uma relação de independência;
- O tamanho da amostra deve ser superior a cerca de sete casos (SIEGEL, 1988, p. 269).

6.4. Resultados obtidos e análise

"A vida é a arte de tirar conclusões suficientes de premissas insuficientes."
[Samuel Butler (*apud* WONNACOTT (1980, p. 487)]

A demonstração bem-sucedida do PROFAX num teste preliminar propiciou o desenvolvimento de um instrumento de avaliação de SS mais aprimorado como o PRONTO© e reuniu os meios necessários para a refutação da hipótese nula.

Foi o intuito inicial para a metodologia de verificação não-paramétrica desta pesquisa aplicar os testes descritos no subitem anterior, segundo a recomendação do *planejamento em blocos* de MOORE (2000, p. 151), que aumentaria o controle sobre o efeito de variáveis ocultas ao experimento, capazes de prejudicar os resultados pela conseqüente tendenciosidade que introduzem.

Sendo assim, três blocos foram originalmente orlados para cada uma das perguntas. A característica fundamental e comum dos três blocos é a formação profissional dos seus componentes (militares da área técnica, especificamente do *geoprocessamento*).

Cada bloco, contudo, seria formado por indivíduos que gozariam de propriedades distintivas, p.ex: um deles (11 respondentes) seria composto por indivíduos da faixa etária média de 25 a 35 anos, mais empregados em funções administrativas (atividade-meio). Outro bloco (42 respondentes) seria composto por indivíduos da faixa etária média entre 20 e 25 anos, mais empregados na atividade-fim, i.e., na área de produção técnica. Um terceiro bloco (14 respondentes) seria composto por indivíduos mais velhos que os dos outros dois blocos (média em torno de 45 anos), constituído de indivíduos com experiência balanceada

(tanto nas técnicas como nas administrativas) e empregados no controle e na coordenação das tarefas executadas pelos indivíduos dos dois primeiros blocos.

Pela premência do prazo de conclusão do trabalho, esse planejamento foi abandonado e todos os indivíduos foram grupados num único bloco de 67 respondentes, com respostas válidas que variavam de 53 a 56 para cada uma das seis perguntas formuladas.

6.4.1. Resultados obtidos

A apresentação dos resultados, a seguir, será realizada por meio de tabelas e gráficos, que sintetizarão, por pergunta (de 1 a 6), a coleta das respostas dos indivíduos e os testes não-paramétricos para a primeira e segunda hipóteses estatísticas, corroborando-as em níveis de significância tolerados para o tipo de manipulação experimental que foi levado a termo nesta pesquisa. .

Como o PRONTO© operou numa região de fronteira entre conhecimentos das ciências sociais (Ciência da Informação e ciências cognitivas) e das formais (Lógica e Computação), é razoável que os índices de validação estatística sejam escolhidos em função do ramo mais complexo (o primeiro), até porque ele é o alvo, a meta, enfim o referencial para ser alcançado pela mediação dos mecanismos formais de explicação e de verificação de teorias, como é o caso do segundo ramo.

Tabela 6.1: Estatísticas descritivas (classes-protótipos) - 1ª Pergunta.

Classes	N	Mínimo	Máximo
Lago	56	1	9
Bosque	56	2	10
Vala	56	2	9
Olho d'água	56	1	9
Caminho carroçável	56	1	10
Estrada-de-ferro	56	1	10
Canal	56	1	8
Cachoeira	56	1	10
Pântano	56	4	10
Riacho	56	1	10

Tendo isso em conta, um nível de significância da ordem de 0,05 é considerado como suficiente por vários autores para experimentos nas áreas das ciências sociais aplicadas e na Psicologia [(DANIEL, 1978), (SIEGEL, 1988) e (MILLER, 2002)]. Como se verá, esse nível de significância correspondeu plenamente aos testes não-paramétricos aplicados. Apesar das limitações experimentadas na metodologia de corroboração das hipóteses, os re-

sultados foram extremamente satisfatórios, ultrapassando o nível de significância de 0,05 (quatro das perguntas chegaram ao nível de 0,01 e uma ao nível de 0,02).

Todos os cálculos que serão apresentados a seguir foram efetuados e ilustrados pelos meios disponíveis no SPSS™ e no PRONTO©.

As tabelas de consolidação dos resultados apurados nos questionários aplicados encontram-se no Apêndice B (CD que acompanha o tomo da tese).

6.4.1.1. Resultados relacionados à primeira pergunta do questionário

A classe variante ou o termo-sujeito correspondente à Tabela 6.1 é RIO.

A Tabela 6.2 indica as médias obtidas para a ordenação (postos) dos julgamentos dos indivíduos para a primeira pergunta.

Tabela 6.2: Média dos postos das respostas dos indivíduos - 1ª Pergunta.

Classes	Médias
Lago	4,79
Bosque	9,02
Vala	4,82
Olho d'água	4,63
Caminho carroçável	7,61
Estrada-de-ferro	8,29
Canal	3,30
Cachoeira	4,55
Pântano	6,50
Riacho	1,50

A Tabela 6.3 apresenta um sumário dos testes não-paramétricos para determinar o grau de relacionamento entre as respostas dos indivíduos ao questionário.

Nessa tabela, a variável "N" representa o número de respostas válidas, "W", o coeficiente de concordância de Kendall e "g.l.", os graus de liberdade (número de variáveis ou de classes –1) associados ao teste quiquadrado.

Tabela 6.3: Resumo dos testes não-paramétricos para as respostas dos indivíduos - 1ª Pergunta.

N	56
W	,594
Quiquadrado	299,540
g.l.	9
Valor de prova	,000

O tamanho da amostra - número de casos ou de respostas válidas dos indivíduos (N = 56) – justifica o emprego do teste quiquadrado.

Tabela 6.4: Sumário dos casos (classes-protótipos) - 1ª Pergunta.

Classes	Médias das Respostas dos Indivíduos	SS do PRONTO	Postos das Respostas dos Indivíduos	Postos dos Resultados do PRONTO
Lago	4,790	,526	5,000	3,000
Bosque	9,020	,044	10,000	9,500
Vala	4,820	,291	6,000	6,000
Olho d'água	4,630	,777	4,000	2,000
Caminho carroçável	7,610	,051	8,000	7,500
Estrada-de-ferro	8,290	,051	9,000	7,500
Canal	3,300	,376	2,000	5,000
Cachoeira	4,550	,416	3,000	4,000
Pântano	6,500	,044	7,000	9,500
Riacho	1,500	1,000	Mais similar	Mais similar

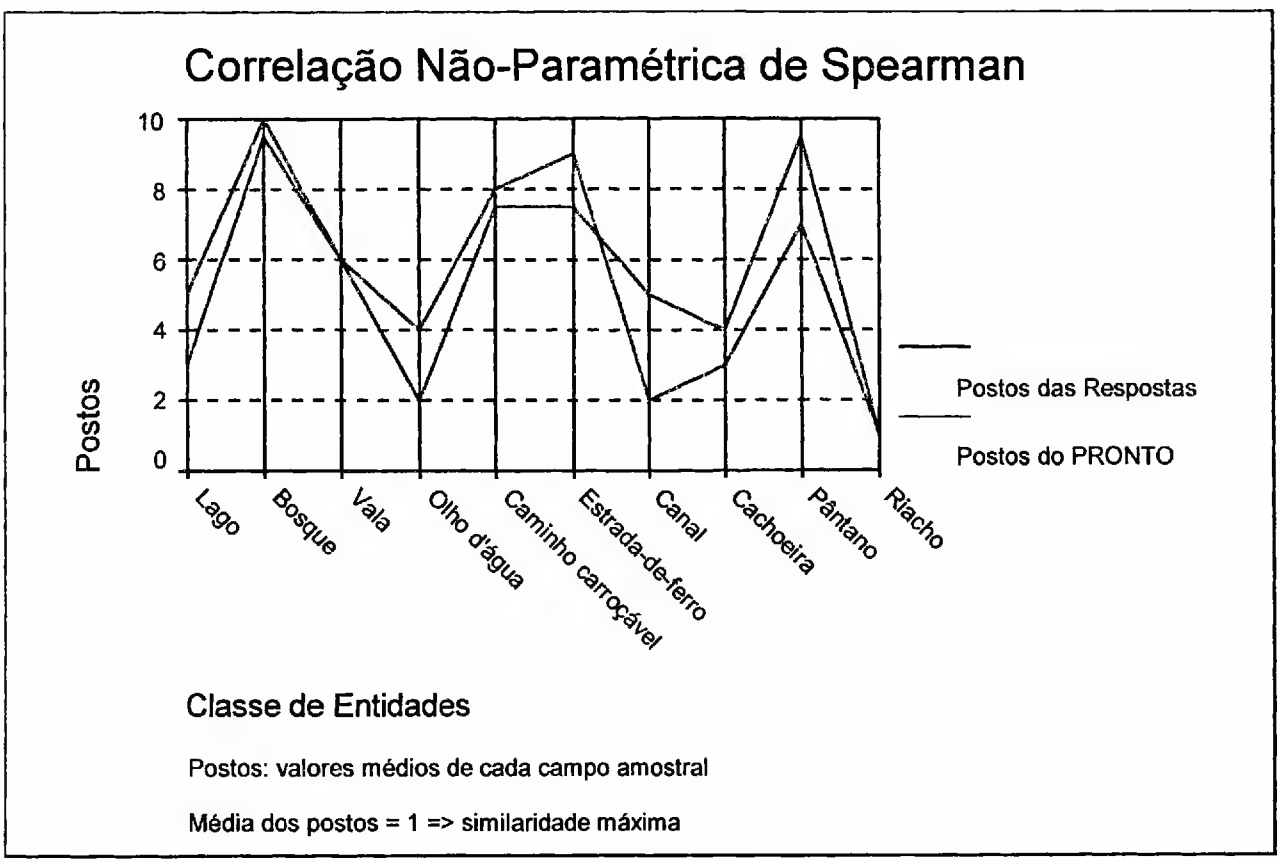

Figura 6.30: Curvas de similaridade das respostas dos indivíduos e dos resultados do PRONTO para a Pergunta 1.

Como o valor de *W* na Tabela 6.3 é de 0,594, é fácil verificar que existe correlação entre as respostas dos indivíduos. A pergunta que se faz, todavia, é a seguinte: Em que grau de confiança pode-se afirmar que as respostas dos indivíduos representam um consenso entre eles?

Pela Equação 6.7, o valor calculado para a estimativa X^2 = 299,54 é bem maior que o valor quiquadrado para um nível de significância crítico de 0,01, com 9 graus de liberdade, que é tabelado em 21,67 (SIEGEL, 1988, p.323). Portanto, como de $X_w^2 >> X^2_{c\,(g.l.=9)}$, o valor de prova é desprezível, comparado com o nível de significância crítico[275], ou melhor, pode-se rejeitar, com baixíssima probabilidade de erro, a hipótese nula de que as respostas dos indivíduos não estejam correlacionadas.

Confirmada a primeira hipótese estatística para a primeira pergunta, é necessário confirmar a segunda hipótese, na qual se afirma que existe uma correlação sensível entre os resultados do PRONTO© e as respostas dos indivíduos, ou melhor, de que o PRONTO© é um instrumento de avaliação de SS capaz de reproduzir o senso humano de julgamento desse fenômeno. Para isso, o coeficiente de concordância de *Kendall* para muitas amostras não é mais o indicado e sim o coeficiente de correlação de *Spearman* (R_s) para duas amostras. A Tabela 6.4 sumaria os valores obtidos para estas duas amostras.

Os termos-protótipos ou predicados representam as dez classes de entidades discriminadas no eixo das abscissas do gráfico da Figura 6.30. Para os indivíduos, a classe menos similar foi BOSQUE. No caso do PRONTO©, para a classe menos similar, houve empate entre BOSQUE e PÂNTANO.

Tabela 6.5: Correlação não-paramétrica de *Spearman* por postos das respostas dos indivíduos e dos resultados do PRONTO- 1ª Pergunta.

		Postos das Respostas	Postos do PRONTO
Postos das Respostas	*Rs*	1,000	,835
	valor de prova	.	,001
	N	10	10
Postos do PRONTO	*Rs*	,835	1,000
	valor de prova	,001	.
	N	10	10

O gráfico da Figura 6.30 mostra uma perfeita coincidência entre os julgamentos humanos de similaridade e os resultados do PRONTO© para os pares formados entre RIO e os seguintes protótipos: VALA e RIACHO. Há também uma grande aproximação para os pares

[275] Mesmo num nível ainda mais rigoroso (de 0,001), ainda continua valendo a rejeição da H_0.

formados entre RIO e os protótipos: BOSQUE, CAMINHO CARROÇÁVEL e CACHOEIRA, o que produziu um elevado valor para R_s.

Finalmente, para a primeira pergunta, a Tabela 6.5 resume o último teste não-paramétrico, que rejeita a hipótese nula e corrobora a segunda hipótese estatística, i.e., corrobora a existência de uma correlação sensível e estatisticamente discernível entre os resultados alcançados pelo PRONTO© e o senso de julgamento de SS dos indivíduos que compararam RIO com as dez classes-protótipos selecionadas.

Por um teste unilateral de hipóteses para *Rs*, num nível de significância (alfa) de 0,02, (nível de confiança maior do que 98%), pode-se rejeitar a hipótese nula, visto que o valor de prova é extremamente menor do que o nível de significância, o que indica haver uma correlação extremamente sensível entre as respostas dos indivíduos e os resultados do protótipo.

Essa conclusão advém da aplicação da Equação 6.6, em que $z_{Rs} = 2,51$. Como o valor crítico da variável aleatória normalizada (z_c), para o nível de significância de 0,02, é tabelado em 2,33 ($z_{Rs} > z_c$), o valor de prova se situa na região de rejeição da hipótese nula.

6.4.1.2. Resultados relacionados à segunda pergunta do questionário

A classe variante ou o termo-sujeito correspondente à Tabela 6.6 é LAGO.

Tabela 6.6: Estatísticas descritivas (classes-protótipos) - 2ª Pergunta.

Classes	N	Mínimo	Máximo
Rio	56	1	10
Pântano	56	1	10
Canal	56	1	8
Cachoeira	56	1	10
Olho d'água	56	1	9
Estrada-de-ferro	56	1	19
Caminho carroçável	56	1	10
Bosque	56	1	10
Riacho	56	1	10
Vala	56	2	10

A Tabela 6.7 indica as médias obtidas para a ordenação (postos) dos julgamentos dos indivíduos para a segunda pergunta.

A Tabela 6.8 apresenta um sumário dos testes não-paramétricos para determinar o grau de relacionamento entre as respostas dos indivíduos ao questionário.

Como o valor de *W* na Tabela 6.8 é de 0,491, é fácil verificar que existe correlação entre as respostas dos indivíduos. A pergunta que se faz, todavia, é a seguinte: Em que grau

de confiança pode-se afirmar que as respostas dos indivíduos representam um consenso entre eles?

Tabela 6.7: Média dos postos das respostas dos indivíduos - 2ª Pergunta.

Classes	Médias
Rio	3,20
Pântano	2,89
Canal	4,71
Cachoeira	4,84
Olho d'água	.4,00
Estrada-de-ferro	9,05
Caminho carroçável	8,39
Bosque	7,25
Riacho	4,63
Vala	6,04

Tabela 6.8: Resumo dos testes não-paramétricos para as respostas dos indivíduos - 2ª Pergunta

N	56
W	**,491**
Quiquadrado	**247,535**
g.l.	9
Valor de prova	,000

Tabela 6.9: Sumário dos casos (classes-protótipos) - 2ª Pergunta.

Classes	Médias das Respostas dos Indivíduos	SS do PRONTO	Postos das Respostas dos Indivíduos	Postos dos Resultados do PRONTO
Rio	3,200	,526	2,000	1,500
Pântano	2,890	,466	Mais similar	5,000
Canal	4,710	,515	5,000	3,000
Cachoeira	4,840	,333	6,000	6,500
Olho d'água	4,000	,333	3,000	6,500
Estrada-de-ferro	9,050	,166	10,000	9,000
Caminho carroçável	8,390	,166	9,000	9,000
Bosque	7,250	,166	8,000	9,000
Riacho	4,630	,526	4,000	1,500
Vala	6,040	,484	7,000	4,000

Pela Equação 6.7, o valor calculado para a estimativa X^2 = 247,535 é bem maior que o valor quiquadrado para um nível de significância crítico de 0,01, com 9 graus de liberdade, que é tabelado em 21,67 (SIEGEL, 1988, p.323). Portanto, como de $X_w^2 >> X^2_{c\ (g.l.=9)}$, o valor de prova é desprezível, comparado com o nível de significância crítico, ou melhor, pode-se

rejeitar, com baixíssima probabilidade de erro, a hipótese nula de que as respostas dos indivíduos não estejam correlacionadas.

Os termos-protótipos ou predicados representam as dez classes de entidades discriminadas no eixo das abscissas do gráfico da Figura 6.31. Para os indivíduos, a classe mais similar foi PÂNTANO, enquanto que para o PRONTO© houve empate entre RIO e RIACHO como classes mais similares a LAGO.

Confirmada a primeira hipótese estatística para a segunda pergunta, é necessário confirmar a segunda hipótese. Para isso, é preciso calcular o coeficiente de correlação de *Spearman* (R_s) para as duas amostras. A Tabela 6.9 sumaria os valores obtidos para essas duas amostras.

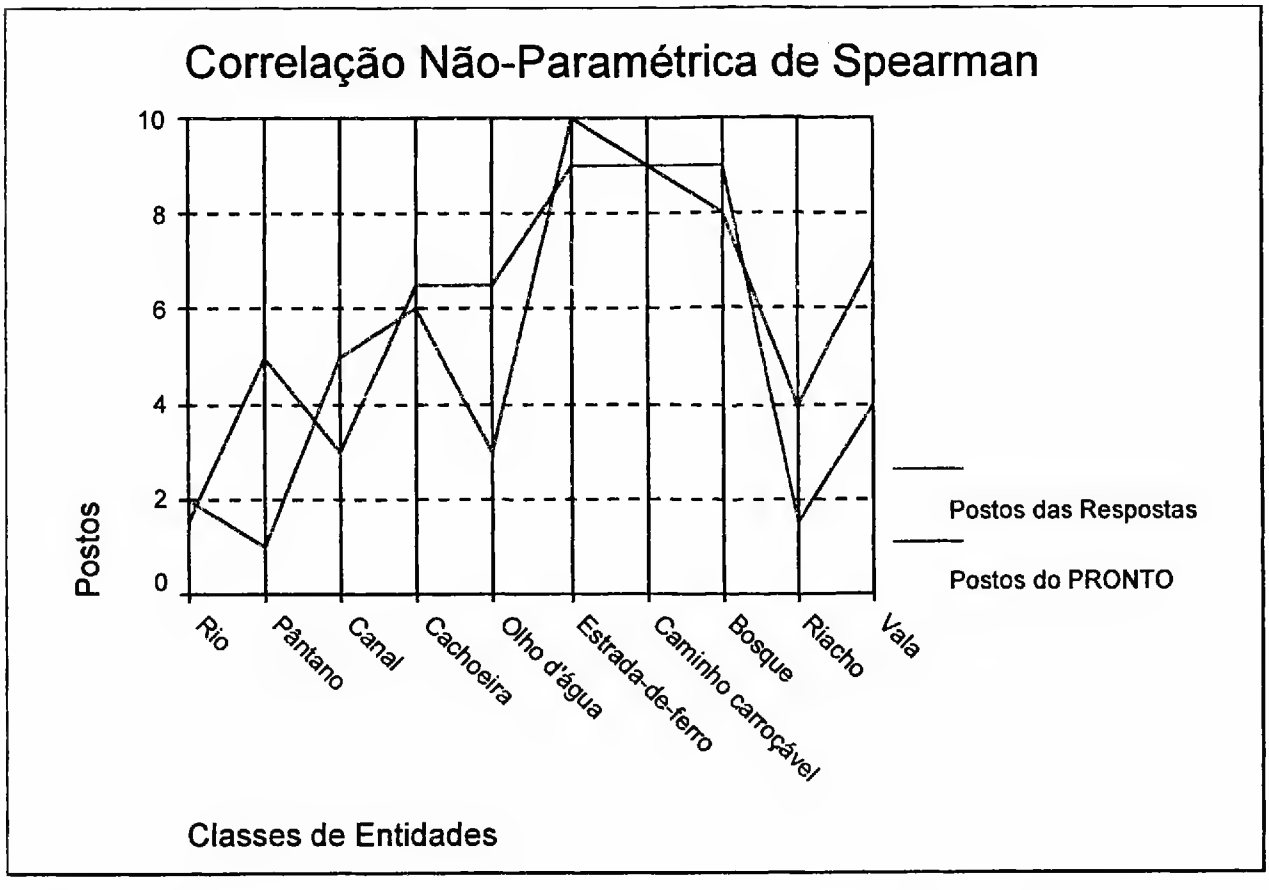

Figura 6.31: Curvas de similaridade das respostas dos indivíduos e dos resultados do PRONTO para a Pergunta 2.

Finalmente, para a segunda pergunta, a Tabela 6.10 resume o último teste não-paramétrico, que rejeita a hipótese nula e corrobora a segunda hipótese estatística, i.e., corrobora a existência de uma correlação sensível e estatisticamente discernível entre os resul-

tados alcançados pelo PRONTO© e o senso de julgamento de SS dos indivíduos que compararam LAGO com as dez classes-protótipos selecionadas.

Tabela 6.10: Correlação não-paramétrica de Spearman por postos das respostas dos indivíduos e dos resultados do PRONTO - 2ª Pergunta.

		Postos das Respostas	Postos do PRONTO
Postos das Respostas	***Rs***	1,000	**,691**
	valor de prova	,	,013
	N	10	10
Postos do PRONTO	***Rs***	**,691**	1,000
	valor de prova	,013	,
	N	10	10

Por um teste unilateral de hipóteses para *Rs*, num nível de significância (alfa) de 0,05 (nível de confiança maior do que 95%), pode-se rejeitar a hipótese nula, visto que o valor de prova é menor do que o nível de significância, o que indica haver uma correlação sensível entre as respostas dos indivíduos e os resultados do protótipo.

Essa conclusão advém da aplicação da Equação 6.6, em que $z_{Rs} = 2,07$. Como o valor crítico da variável aleatória normalizada (z_c) para o nível de significância de 0,05 é tabelado em 1,96 ($z_{Rs} > z_c$), o valor de prova se situa na região de rejeição da hipótese nula.

6.4.1.3. Resultados relacionados à terceira pergunta do questionário

A classe variante ou o termo-sujeito correspondente à Tabela 6.11 é ÁREA DE LAZER.

Tabela 6.11: Estatísticas descritivas (classes-protótipos) - 3ª Pergunta.

Classes	N	Mínimo	Máximo
Campus universitário	56	1	10
Represa	56	2	10
Campo de futebol	56	1	8
Praça de esportes	56	1	10
Campo de tiro-ao-alvo	56	2	10
Estádio	56	1	9
Museu	56	2	9
Estação ferroviária	56	1	10
Vila	56	4	10
Piscina	56	1	10

A Tabela 6.12 indica as médias obtidas para a ordenação (postos) dos julgamentos dos indivíduos para a terceira pergunta.

A Tabela 6.13 apresenta um sumário dos testes não-paramétricos para determinar o grau de relacionamento entre as respostas dos indivíduos ao questionário.

Tabela 6.12: Média dos postos das respostas dos indivíduos - 3ª Pergunta.

Classes	Médias
Campus universitário	6,79
Represa	8,00
Campo de futebol	3,16
Praça de esportes	1,63
Campo de tiro-ao-alvo	5,45
Estádio	3,82
Museu	5,43
Estação ferroviária	8,52
Vila	7,82
Piscina	4,39

Como o valor de *W* na Tabela 6.13 é de 0,569, é fácil verificar que existe correlação entre as respostas dos indivíduos. A pergunta que se faz, todavia, é a seguinte: Em que grau de confiança pode-se afirmar que as respostas dos indivíduos representam um consenso entre eles?

Tabela 6.13: Resumo dos testes não-paramétricos para as respostas dos indivíduos - 3ª Pergunta.

N	56
W	**,569**
Quiquadrado	**286,753**
g.l.	9
Valor de prova	,000

Tabela 6.14: Sumário dos casos (classes-protótipos) - 3ª Pergunta.

Classes	Médias das Respostas dos Indivíduos	SS do PRONTO	Postos das Respostas dos Indivíduos	Postos dos Resultados do PRONTO
Campus universitário	6,790	,054	7,000	9,000
Represa	8,000	,085	9,000	8,000
Campo de futebol	3,160	,484	2,000	3,000
Praça de esportes	1,630	,789	Mais similar	Mais similar
Campo de tiro-ao-alvo	5,450	,333	6,000	5,000
Estádio	3,820	,526	3,000	2,000
Museu	5,430	,222	5,000	6,000
Estação ferroviária	8,520	,000	10,000	10,000
Vila	7,820	,166	8,000	7,000
Piscina	4,390	,356	4,000	4,000

Pela Equação 6.7, o valor calculado para a estimativa X^2 = 286,753 é bem maior que o valor quiquadrado para um nível de significância crítico de 0,01, com 9 graus de liberdade, que é tabelado em 21,67 (SIEGEL, 1988, p.323). Portanto, como de $X_w^2 >> X^2_{c\,(g.l.=9)}$, o valor

de prova é desprezível, comparado com o nível de significância crítico, ou melhor, pode-se rejeitar, com baixíssima probabilidade de erro, a hipótese nula de que as respostas dos indivíduos não estejam correlacionadas.

Confirmada a primeira hipótese estatística para a terceira pergunta, é necessário confirmar a segunda hipótese. Para isso, é preciso calcular o coeficiente de correlação de *Spearman* (R_s) para as duas amostras. A Tabela 6.14 sumaria os valores obtidos para estas duas amostras.

Os termos-protótipos ou predicados representam as dez classes de entidades discriminadas no eixo das abscissas do gráfico da Figura 6.32. Para os indivíduos e para o protótipo, a classe menos similar em relação à classe ÁREA DE LAZER (termo variante) foi ESTAÇÃO FERROVIÁRIA.

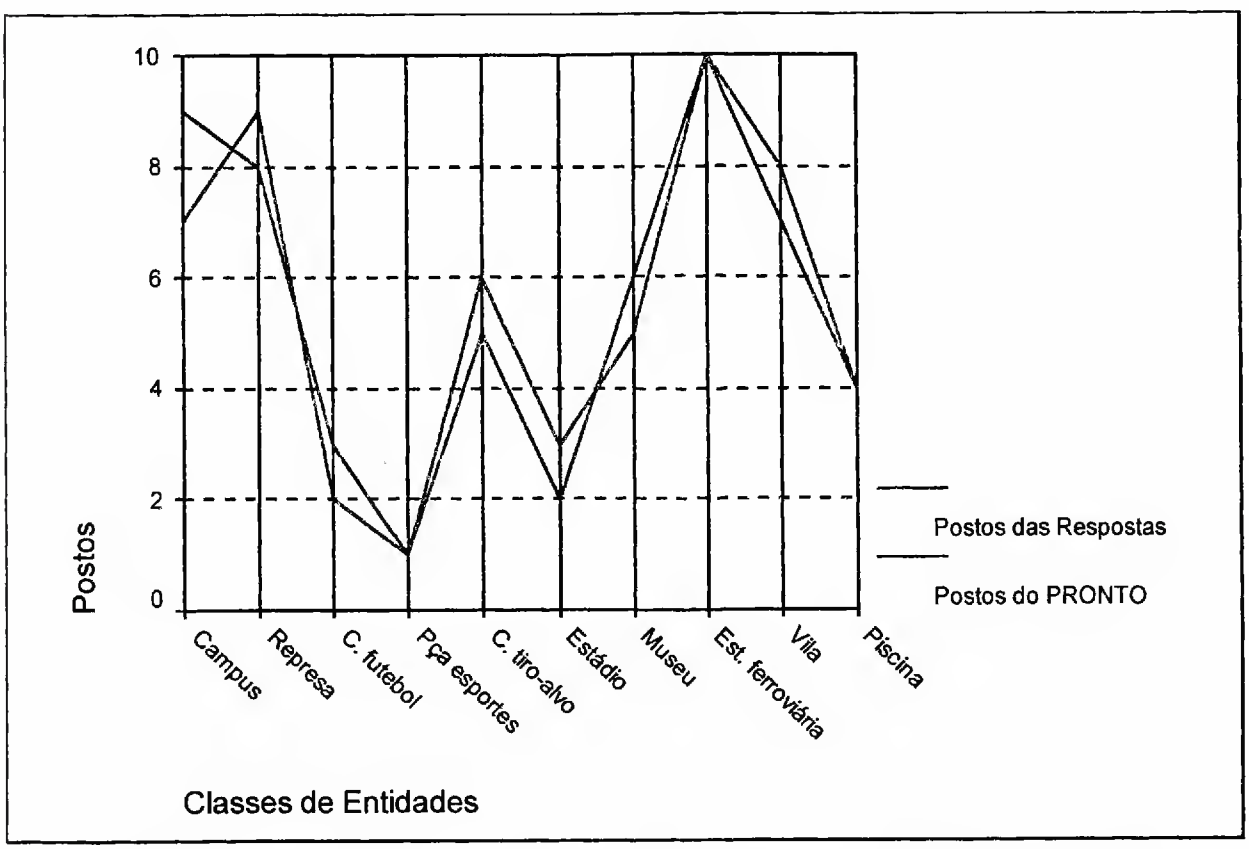

Figura 6.32: Curvas de similaridade das respostas dos indivíduos e dos resultados do PRONTO para a Pergunta 3.

O gráfico mostra uma perfeita coincidência entre os julgamentos humanos de similaridade e os resultados do PRONTO© para os pares formados por ÁREA DE LAZER com os protótipos: PRAÇA DE ESPORTES e ESTAÇÃO FERROVIÁRIA (V. subitem 6.4.2).

Tabela 6.15: Correlação não-paramétrica de Spearman por postos das respostas dos indivíduos e dos resultados do PRONTO - 3ª Pergunta.

		Postos das Respostas	Postos do PRONTO
Postos das Respostas	*Rs*	1,000	**,891**
	valor de prova	.	,000
	N	10	10
Postos do PRONTO	*Rs*	**,891**	1,000
	valor de prova	,000	.
	N	10	10

Finalmente, para a terceira pergunta, a Tabela 6.15 resume o último teste não-paramétrico, que rejeita a hipótese nula e corrobora a segunda hipótese estatística, i.e., corrobora a existência de uma correlação sensível e estatisticamente discernível entre os resultados alcançados pelo PRONTO© e o senso de julgamento de SS dos indivíduos que compararam ÁREA DE LAZER com as dez classes-protótipos selecionadas.

Por um teste unilateral de hipóteses para *Rs*, num nível de significância (alfa) de 0,01 (nível de confiança maior do que 99%), pode-se rejeitar a hipótese nula, visto que o valor de prova é extremamente menor do que o nível de significância, o que indica haver uma correlação sensível entre as respostas dos indivíduos e os resultados do protótipo.

Tabela 6.16: Estatísticas descritivas (classes-protótipos) - 4ª Pergunta.

Classes	N	Mínimo	Máximo
Aeroporto	55	1	10
Pântano	55	2	10
Armazém	55	1	10
Atracadouro	55	1	10
Recife	55	1	10
Estrada-de-ferro	55	2	10
Lago	55	2	10
Rio	55	1	10
Pátio ferroviário	55	1	9
Morro	55	2	10

Essa conclusão advém da aplicação da Equação 6.6, em que $z_{Rs} = 2,67$. Como o valor crítico da variável aleatória normalizada (z_c) para o nível de significância de 0,01 é tabelado em 2,57 ($z_{Rs} > z_c$), o valor de prova se situa na região de rejeição da hipótese nula.

6.4.1.4. Resultados relacionados à quarta pergunta do questionário

A classe variante ou o termo-sujeito correspondente à Tabela 6.16 é CAIS.

Tabela 6.17: Média dos postos das respostas dos indivíduos - 4ª Pergunta.

Classes	Médias
Aeroporto	4,80
Pântano	7,38
Armazém	4,55
Atracadouro	1,71
Recife	6,00
Estrada-de-ferro	6,36
Lago	5,82
Rio	5,40
Pátio ferroviário	4,33
Morro	8,65

A Tabela 6.17 indica as médias obtidas para a ordenação (postos) dos julgamentos dos indivíduos para a quarta pergunta.

A Tabela 6.18 apresenta um sumário dos testes não-paramétricos para determinar o grau de relacionamento entre as respostas dos indivíduos ao questionário.

Como o valor de *W* na Tabela 6.18 é de 0,385, é fácil verificar que existe correlação entre as respostas dos indivíduos. A pergunta que se faz, todavia, é a seguinte: Em que grau de confiança pode-se afirmar que as respostas dos indivíduos representam um consenso entre eles?

Tabela 6.18: Resumo dos testes não-paramétricos para as respostas dos indivíduos - 4ª Pergunta.

N	55
W	**,385**
Quiquadrado	**190,482**
g.l.	9
Valor de prova	,000

Pela Equação 6.7, o valor calculado para a estimativa X^2 = 190,482 ainda é bem maior que o valor quiquadrado para um nível de significância crítico de 0,01, com 9 graus de liberdade, que é tabelado em 21,67 (SIEGEL, 1988, p.323). Portanto, como de $X_w^2 >> X^2_{c\,(g.l.=9)}$, o valor de prova é desprezível, comparado com o nível de significância crítico, ou melhor, pode-se rejeitar, com baixíssima probabilidade de erro, a hipótese nula de que as respostas dos indivíduos não estejam correlacionadas.

Confirmada a primeira hipótese estatística para a quarta pergunta, é necessário confirmar a segunda hipótese. Para isso, é preciso calcular o coeficiente de correlação de *Spe-*

arman (R_s) para as duas amostras. A Tabela 6.19 sumaria os valores obtidos para estas duas amostras.

Tabela 6.19: Sumário dos casos (classes-protótipos) - 4ª Pergunta.

Classes	Médias das Respostas dos Indivíduos	SS do PRONTO	Postos das Respostas dos Indivíduos	Postos dos Resultados do PRONTO
Aeroporto	4,800	,615	4,000	2,000
Pântano	7,380	,166	9,000	7,000
Armazém	4,550	,484	3,000	4,000
Atracadouro	1,710	,589	Mais similar	3,000
Recife	6,000	,001	7,000	9,000
Estrada-de-ferro	6,360	,085	8,000	8,000
Lago	5,820	,333	6,000	5,000
Rio	5,400	,222	5,000	6,000
Pátio ferroviário	4,330	,777	2,000	Mais similar
Morro	8,650	,000	10,000	10,000

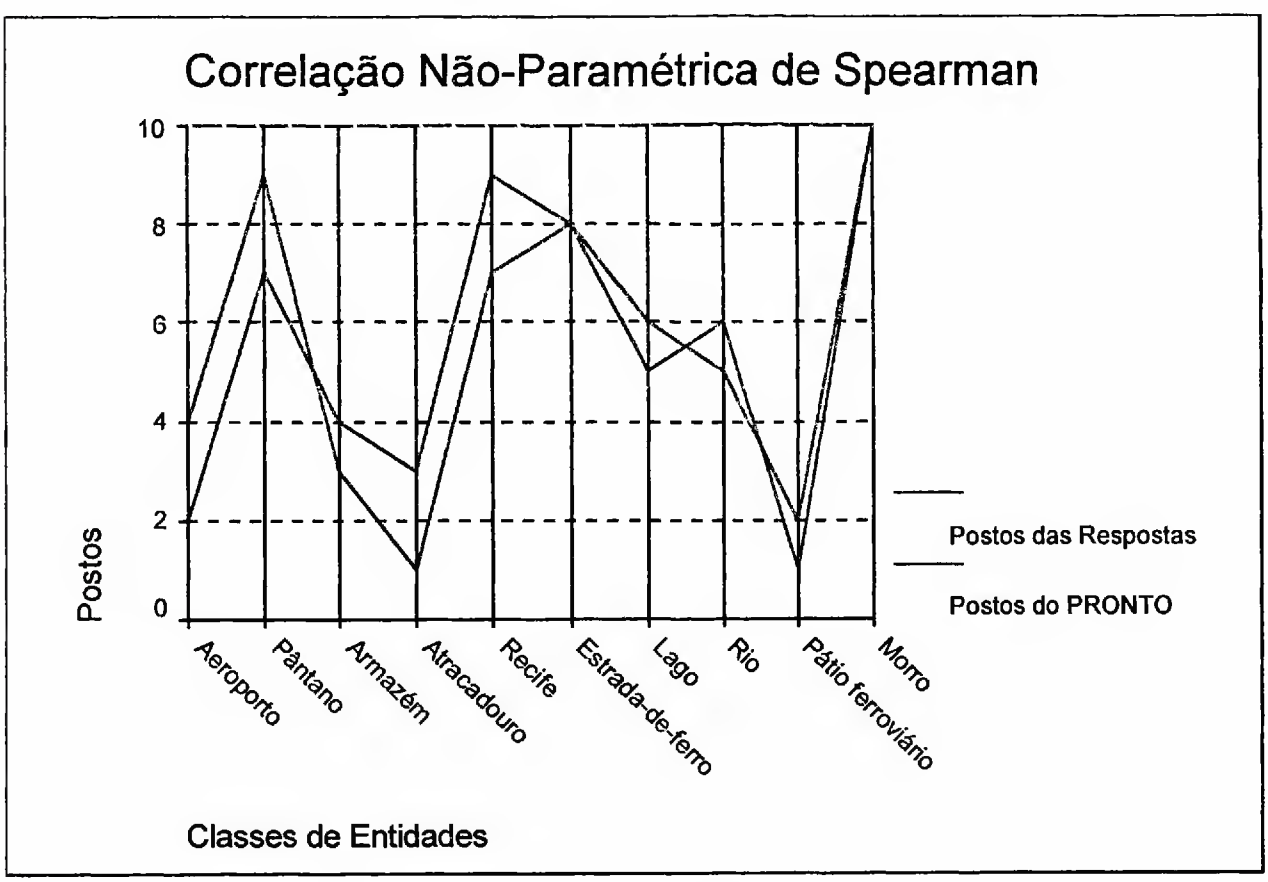

Figura 6.33: Curvas de similaridade das respostas dos indivíduos e dos resultados do PRONTO para a Pergunta 4.

Os termos-protótipos ou predicados representam as dez classes de entidades discriminadas no eixo das abscissas do gráfico da Figura 6.33. Para os indivíduos, a classe mais similar em relação à classe CAIS (termo variante) foi ATRACADOURO. Para o protótipo, foi PÁTIO FERROVIÁRIO. A menos similar para os indivíduos e para o protótipo foi MORRO.

O gráfico mostra uma perfeita coincidência entre os julgamentos humanos de similaridade e os resultados do PRONTO© para os pares formados por CAIS com os protótipos: MORRO e ESTRADA-DE-FERRO. Há também uma grande aproximação para os pares formados entre CAIS e os protótipos: RIO, LAGO e ARMAZÉM, o que produziu um elevado valor para *Rs*.

Finalmente, para a quarta pergunta, a Tabela 6.20 resume o último teste não-paramétrico, que rejeita a hipótese nula e corrobora a segunda hipótese estatística, i.e., corrobora a existência de uma correlação sensível e estatisticamente discernível entre os resultados alcançados pelo PRONTO© e o senso de julgamento de SS dos indivíduos que compararam CAIS com as dez classes-protótipos selecionadas.

Por um teste unilateral de hipóteses para *Rs*, num nível de significância (alfa) de 0,01 (nível de confiança maior do que 99%), pode-se rejeitar a hipótese nula, visto que o valor de prova é extremamente menor do que o nível de significância, o que indica haver uma correlação sensível entre as respostas dos indivíduos e os resultados do protótipo.

Essa conclusão advém da aplicação da Equação 6.6, em que $z_{Rs} = 2,64$. Como o valor crítico da variável aleatória normalizada (z_c) para o nível de significância de 0,01 é tabelado em 2,57 ($z_{Rs} > z_c$), o valor de prova se situa na região de rejeição da hipótese nula.

Tabela 6.20: Correlação não-paramétrica de Spearman por postos das respostas dos indivíduos e dos resultados do PRONTO - 4ª Pergunta.

		Postos das Respostas	Postos do PRONTO
Postos das Respostas	*Rs*	1,000	**,879**
	valor de prova	,	,000
	N	10	10
Postos do PRONTO	*Rs*	**,879**	1,000
	valor de prova	,000	,
	N	10	10

6.4.1.5. Resultados relacionados à quinta pergunta do questionário

A classe variante ou o termo-sujeito correspondente à Tabela 6.21 é CANAL.

Tabela 6.21: Estatísticas descritivas (classes-protótipos) - 5ª Pergunta.

Classes	N	Mínimo	Máximo
Riacho	53	1	10
Vala	53	1	10
Bosque	53	3	10
Olho d'água	53	1	10
Caminho carroçável	53	1	9
Estrada-de-ferro	53	1	10
Cachoeira	53	2	9
Pântano	53	2	10
Lago	53	2	10

A Tabela 6.22 indica as médias obtidas para a ordenação (postos) dos julgamentos dos indivíduos para a quinta pergunta.

Tabela 6.22: Média dos postos das respostas dos indivíduos - 5ª Pergunta.

Classes	Médias
Riacho	2,45
Vala	2,11
Bosque	7,68
Olho d'água	4,81
Caminho carroçável	5,75
Estrada-de-ferro	6,45
Cachoeira	4,83
Pântano	5,94
Lago	4,96

Tabela 6.23: Resumo dos testes não-paramétricos para as respostas dos indivíduos - 5ª Pergunta.

N	53
W	**,427**
Quiquadrado	**181,162**
g.l.	8
Valor de prova	,000

Confirmada a primeira hipótese estatística para a quinta pergunta, é necessário confirmar a segunda hipótese. Para isso, é preciso calcular o coeficiente de correlação de *Spearman* (R_s) para as duas amostras. A Tabela 6.24 sumaria os valores obtidos para estas duas amostras.

Como o valor de W na Tabela 6.23 é de 0,427, é fácil verificar que existe correlação entre as respostas dos indivíduos. A pergunta que se faz, todavia, é a seguinte: Em que grau

de confiança pode-se afirmar que as respostas dos indivíduos representam um consenso entre eles?

Tabela 6.24: Sumário dos casos (classes-protótipos) - 5ª Pergunta.

Classes	Médias das Respostas dos Indivíduos	SS do PRONTO	Postos das Respostas dos Indivíduos	Postos dos Resultados do PRONTO
Riacho	2,450	,376	2,000	3,000
Vala	2,110	,762	Mais similar	Mais similar
Bosque	7,680	,000	9,000	8,500
Olho d'água	4,810	,333	3,000	4,500
Caminho carroçável	5,750	,111	6,000	6,500
Estrada-de-ferro	6,450	,111	8,000	6,500
Cachoeira	4,830	,333	4,000	4,500
Pântano	5,940	,000	7,000	8,500
Lago	4,960	,615	5,000	2,000

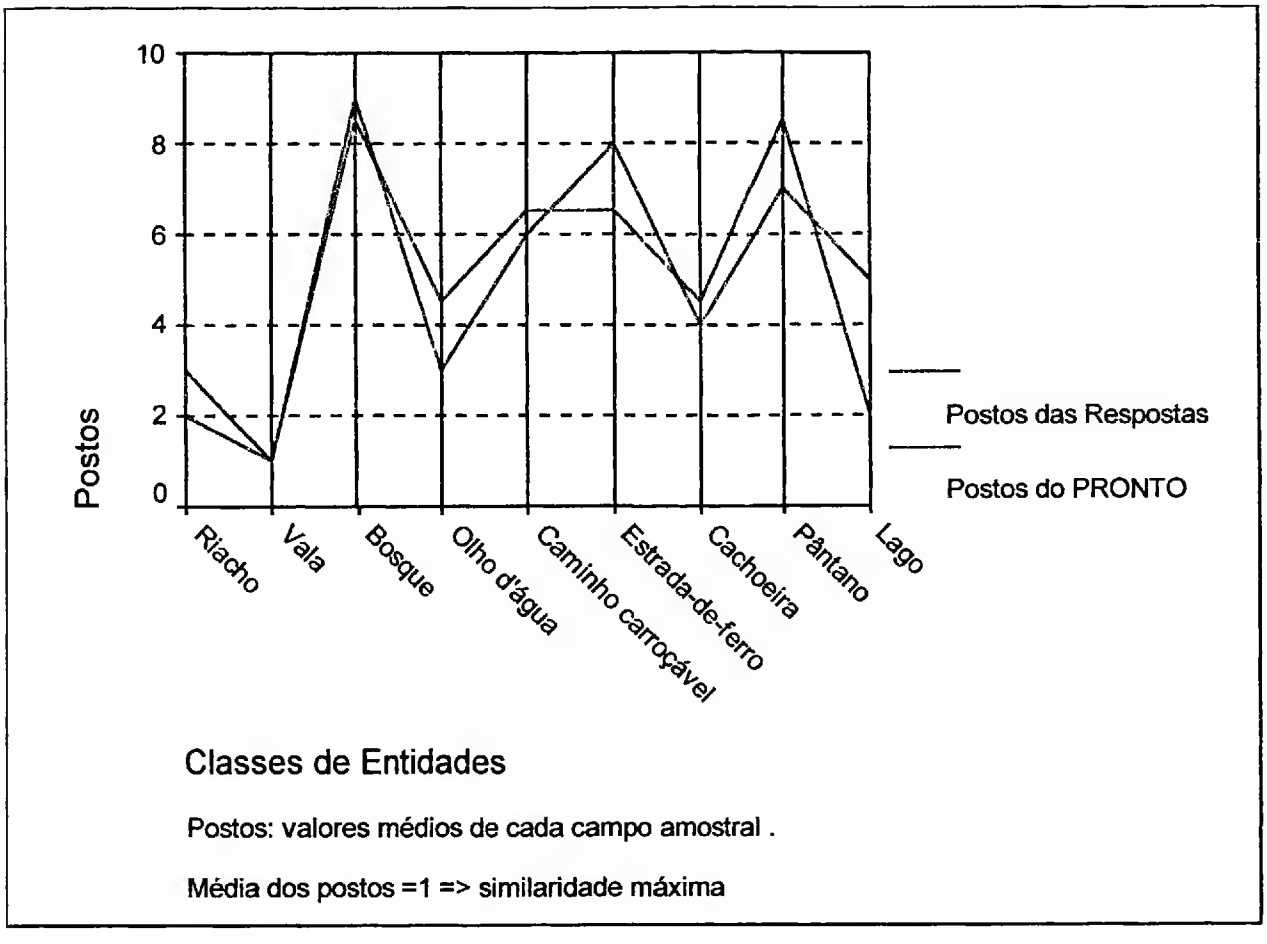

Figura 6.34: Curvas de similaridade das respostas dos indivíduos e dos resultados do PRONTO para a Pergunta 5.

Pela Equação 6.7, o valor calculado para a estimativa X^2 = 181,162 é bem maior que o valor quiquadrado para um nível de significância crítico de 0,01, com 8 graus de liberdade, que é tabelado em 20,09 (SIEGEL, 1988, p.323). Portanto, como de X_w^2 >> $X^2_{c\ (g.l.=9)}$, o valor de prova é desprezível, comparado com o nível de significância crítico, ou melhor, pode-se rejeitar, com baixíssima probabilidade de erro, a hipótese nula de que as respostas dos indivíduos não estejam correlacionadas.

Os termos-protótipos ou predicados representam as dez classes de entidades discriminadas no eixo das abscissas do gráfico da Figura 6.33. Para os indivíduos e para o protótipo, a classe mais similar em relação à classe CANAL (termo variante) foi VALA. A menos similar para os indivíduos e para o protótipo foi BOSQUE, mas o protótipo também colocou empatada, nessa situação, a classe PÂNTANO.

Além das coincidências acima, o gráfico mostra grandes aproximações para os pares formados entre CANAL e os protótipos: RIACHO, OLHO D'ÁGUA, CAMINHO CARROÇÁVEL e CACHOEIRA, o que produziu um elevado valor para *Rs*.

Tabela 6.25: Correlação não-paramétrica de Spearman por postos das respostas dos indivíduos e dos resultados do PRONTO - 5ª Pergunta.

		Postos das Respostas	Postos do PRONTO
Postos das Respostas	*Rs*	1,000	**,852**
	valor de prova	,	,002
	N	9	9
Postos do PRONTO	*Rs*	**,852**	1,000
	valor de prova	,002	,
	N	9	9

Finalmente, para a quinta pergunta, a Tabela 6.25 resume o último teste não-paramétrico, que rejeita a hipótese nula e corrobora a segunda hipótese estatística, i.e., corrobora a existência de uma correlação sensível e estatisticamente discernível entre os resultados alcançados pelo PRONTO© e o senso de julgamento de SS dos indivíduos que compararam CANAL com as dez classes-protótipos selecionadas.

Por um teste unilateral de hipóteses para *Rs*, num nível de significância (alfa) de 0,02 (nível de confiança maior do que 98%), pode-se rejeitar a hipótese nula, visto que o valor de prova é extremamente menor do que o nível de significância, o que indica haver uma correlação sensível entre as respostas dos indivíduos e os resultados do protótipo.

Essa conclusão advém da aplicação da Equação 6.6, em que z_{Rs} = 2,43. Como o valor crítico da variável aleatória normalizada (z_c) para o nível de significância de 0,02 é tabelado em 2,33 ($z_{Rs} > z_c$), o valor de prova se situa na região de rejeição da hipótese nula.

No caso em tela, é de bom alvitre registrar que houve uma falha na aplicação do questionário que, no entanto, não afetou os resultados, porque foram escoimadas as observações redundantes (*outliers* ou *missing* values) referentes à comparação do termo-sujeito (CANAL) com ele mesmo, mas na posição de predicado ou protótipo.

6.4.1.6. Resultados relacionados à sexta pergunta do questionário

A classe variante ou o termo-sujeito correspondente à Tabela 6.26 é ESTÁDIO.

Tabela 6.26: Estatísticas descritivas (classes-protótipos) - 6ª Pergunta.

Classes	N	Mínimo	Máximo
Represa	55	2	10
Campus universitário	55	1	10
Piscina	55	2	10
Praça de esportes	55	1	10
Campo de tiro-ao-alvo	55	2	10
Museu	55	1	10
Estação ferroviária	55	1	10
Vila	55	1	10
Campo de futebol	55	1	10

A Tabela 6.27 indica as médias obtidas para a ordenação (postos) dos julgamentos dos indivíduos para a sexta pergunta.

Tabela 6.27: Média dos postos das respostas dos indivíduos - 6ª Pergunta.

Classes	Médias
Represa	7,15
Campus universitário	4,93
Piscina	4,75
Praça de esportes	2,44
Campo de tiro-ao-alvo	4,35
Museu	6,02
Estação ferroviária	6,84
Vila	6,33
Campo de futebol	2,22

A Tabela 6.28 apresenta um sumário dos testes não-paramétricos para determinar o grau de relacionamento entre as respostas dos indivíduos ao questionário.

Tabela 6.28: Resumo dos testes não-paramétricos para as respostas dos indivíduos - 6ª Pergunta.

N	55
W	**,426**
Quiquadrado	**187,607**
g.l.	8
Valor de prova	,000

Como o valor de *W* na Tabela 6.28 é de 0,426, é fácil verificar que existe correlação entre as respostas dos indivíduos. A pergunta que se faz, todavia, é a seguinte: „Em que grau de confiança pode-se afirmar que as respostas dos indivíduos representam um consenso entre eles?

Tabela 6.29: Sumário dos casos (classes-protótipos) - 6ª Pergunta.

Classes	Médias das Respostas dos Indivíduos	SS do PRONTO	Postos das Respostas dos Indivíduos	Postos dos Resultados do PRONTO
Represa	7,150	,054	9,000	8,000
Campus universitário	4,930	,222	5,000	5,000
Piscina	4,750	,526	4,000	3,000
Praça de esportes	2,440	,888	2,000	Mais similar
Campo de tiro-alvo	4,350	,388	3,000	4,000
Museu	6,020	,085	6,000	7,000
Estação ferroviária	6,840	,001	8,000	9,000
Vila	6,330	,166	7,000	6,000
Campo de futebol	2,220	,762	Mais similar	2,000

Pela Equação 6.7, o valor calculado para a estimativa X^2 = 187,607 é bem maior que o valor quiquadrado para um nível de significância crítico de 0,01, com 8 graus de liberdade, que é tabelado em 20,09 (SIEGEL, 1988, p.323). Portanto, como de $X_w^2 >> X^2_{c\,(g.l.=9)}$, o valor de prova é desprezível, comparado com o nível de significância crítico, ou melhor, pode-se rejeitar, com baixíssima probabilidade de erro, a hipótese nula de que as respostas dos indivíduos não estejam correlacionadas.

Confirmada a primeira hipótese estatística para a sexta pergunta, é necessário confirmar a segunda hipótese. Para isso, é preciso calcular o coeficiente de correlação de *Spearman* (R_s) para as duas amostras. A Tabela 6.29 sumaria os valores obtidos para essas duas amostras.

Os termos-protótipos ou predicados representam as dez classes de entidades discriminadas no eixo das abscissas do gráfico da Figura 6.35. Para os indivíduos, a classe mais similar a ESTÁDIO foi CAMPO DE FUTEBOL. Para o protótipo, foi PRAÇA DE ESPORTES.

A menos similar para os indivíduos foi REPRESA. Para o protótipo, foi ESTAÇÃO FERROVIÁRIA.

Além das coincidências acima, ainda ocorreu mais uma entre ESTÁDIO e *CAMPUS* UNIVERSITÁRIO, para ambos os campos de observação (indivíduos e protótipo).

O gráfico ainda mostra grandes aproximações para os pares formados entre ESTÁDIO e os protótipos: REPRESA, PISCINA, CAMPO DE TIRO-AO-ALVO, MUSEU E VILA, o que produziu um elevado valor para *Rs*.

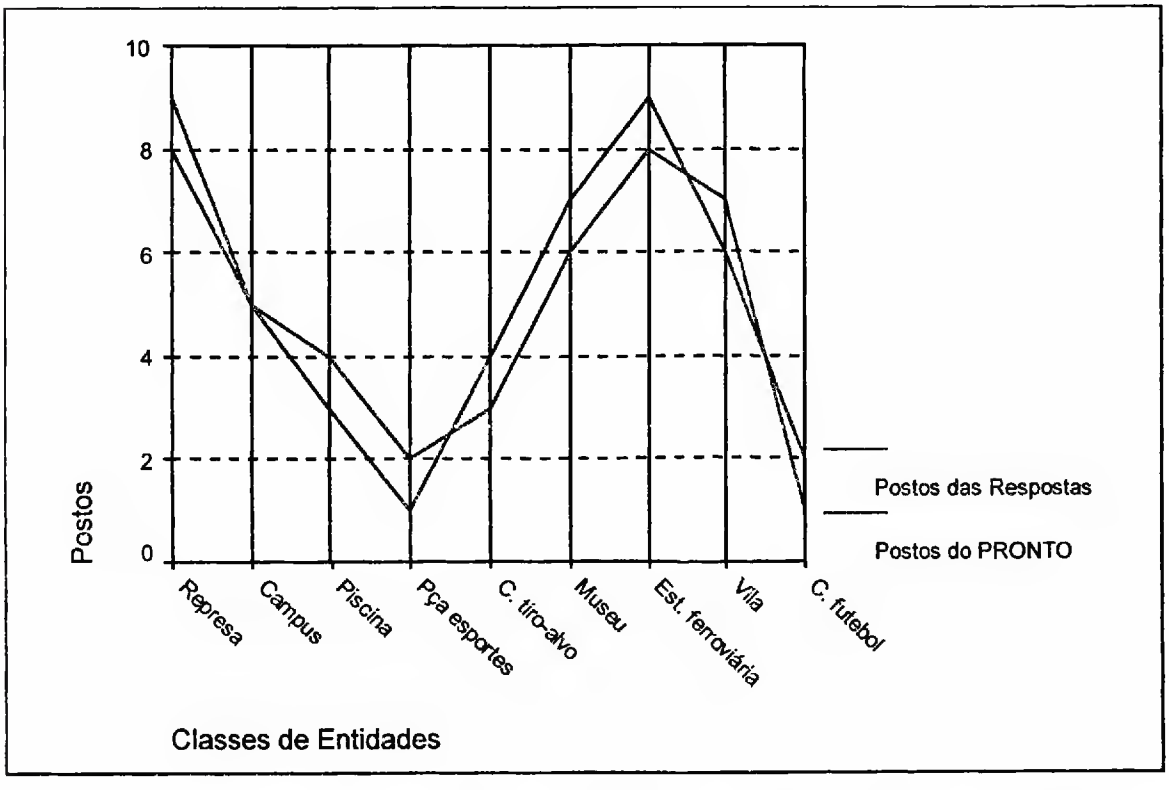

Figura 6.35: Curvas de similaridade das respostas dos indivíduos e dos resultados do PRONTO para a Pergunta 6.

Finalmente, para a sexta pergunta, a Tabela 6.30 resume o último teste não-paramétrico, que rejeita a hipótese nula e corrobora a segunda hipótese estatística, i.e., corrobora a existência de uma correlação sensível e estatisticamente discernível entre os resultados alcançados pelo PRONTO© e o senso de julgamento de SS dos indivíduos que compararam CANAL com as dez classes-protótipos selecionadas.

Por um teste unilateral de hipóteses para *Rs*, num nível de significância (alfa) de 0,01 (nível de confiança maior do que 99%), pode-se rejeitar a hipótese nula, visto que o valor de

prova é extremamente menor do que o nível de significância, o que indica haver uma correlação sensível entre as respostas dos indivíduos e os resultados do protótipo.

Essa conclusão advém da aplicação da Equação 6.6, em que $z_{Rs} = 2,64$. Como o valor crítico da variável aleatória normalizada (z_c) para o nível de significância de 0,01 é tabelado em 2,57($z_{Rs} > z_c$), o valor de prova se situa na região de rejeição da hipótese nula.

Tabela 6.30: Correlação não-paramétrica de Spearman por postos das respostas dos indivíduos e dos resultados do PRONTO - 6ª Pergunta.

		Postos das Respostas	Postos do PRONTO
Postos das Respostas	*Rs*	1,000	**,933**
	valor de prova	,	,000
	N	9	9
Postos do PRONTO	*Rs*	**,933**	1,000
	valor de prova	,000	,
	N	9	9

No caso em tela, é de bom alvitre registrar que houve uma falha na aplicação do questionário que, no entanto, não afetou os resultados, porque foram escoimadas as observações redundantes (*outliers* ou *missing* values) referentes à comparação do termo-sujeito (ESTÁDIO) com ele mesmo, mas na posição de predicado ou protótipo.

6.4.2. Análise dos resultados obtidos

Quanto à metodologia estatística empregada para corroborar a hipótese alternativa desta pesquisa, cabe levantar algumas considerações de ordem estocástica, colocadas por LEVIN (1987, p. 202) e WONNACOTT (1980, p. 387), que frisaram não ser plausível inferir de uma correlação a existência de uma relação obrigatória de **causa e efeito** entre as variáveis testadas pelos métodos não-paramétricos. O que existe é uma **indicação** de que as variáveis verificadas "caminham juntas" - ou em sentidos opostos - e de uma forma mais forte ou não. A correlação constitui o núcleo metodológico deste estudo-de-caso exploratório, do tipo *manipulação experimental*.

Os coeficientes não-paramétricos adotados na manipulação experimental deste estudo-de-caso não têm o poder de estabelecer relações de causa e efeito entre a avaliação de SS realizada pelo protótipo e os fatores de ordem cognitiva que levaram os indivíduos a responder os questionários de uma forma ou de outra. Percebe-se, aí, a necessidade de uma análise de cunho qualitativo, para complementar a consistência das hipóteses, fornecida pelo método quantitativo empregado.

Para estabelecer um teste mais rigoroso, que determinasse relações de causa e efeito entre os julgamentos humanos de SS e os resultados do PRONTO©, a pesquisa não poderia ser um caso especial de pesquisa exploratória, mas sim uma pesquisa experimental, propriamente dita, que nem se coaduna com os objetivos aqui estabelecidos.

No entanto, o estudo de correlação aqui empreendido contribuiu para aumentar o entendimento do fenômeno da SS, dentro das limitações desta pesquisa., bem como indicou possíveis fatores causais que podem ser posteriormente verificados em estudos futuros mais rigorosos (experimentais).

Eis por que não se concluiu com rigor determinístico sobre alguns dos resultados obtidos. A razão disso é que o fenômeno da SS possui um importante componente de natureza psicológica, que não pode ser reduzido facilmente a modelos matemáticos.

Dessa forma, procurou-se complementar uma metodologia quantitativa com outra qualitativa em cada fase do trabalho – no planejamento do experimento, na coleta de dados e na análise dos resultados.

Os mandamentos ortodoxos de metodologia para ambas as modalidades (quantitativa e qualitativa), nas duas primeiras fases de trabalho (plano do experimento e coleta de dados), a bem da honestidade científica, não foram estritamente seguidos, fundamentalmente em virtude do apertado concedido pela organização de origem do pesquisador.

Apesar de desnecessário este registro, já que os resultados quantitativos mostraram-se satisfatórios e as análises qualitativas que se seguem mostrarem-se coerentes com tais resultados, além de estarem fundadas na experiência do pesquisador sobre parte (a cartográfica) do domínio do problema , é de bom alvitre excluir a negligência premeditada, relatando as dificuldades sofridas e os meios de fortuna encontrados para eliminá-las ou, pelo menos, para contorná-las.

Sendo assim, os seguintes procedimentos tradicionais da metodologia não ocorreram ou foram fracamente executados:

- Discussão prévia com o grupo que participou da investigação (Nota: somente os indivíduos que exerciam cargos de chefia - 7 de 67 - foram postos a par das particularidades da pesquisa);
- Não houve a oportunidade (cronológica e material) de se aplicar um questionário prévio para um subgrupo amostral.

O relaxamento perpetrado teve reflexos negativos nas seguintes etapas do trabalho:

- Na formulação dos tópicos operacionais da pesquisa (objetivos específicos e hipóteses estatísticas), tornando árdua a tarefa de concluí-los;

- O questionário prévio poderia ensejar dúvidas no subgrupo do pré-teste, que serviriam para corrigir o rumo do questionário definitivo (Nota: ocorreram 3 equívocos na elaboração do questionário único, que foram compensados como se explicará a seguir);
- O questionário prévio também contribuiria para uma identificação mais adequada de classes de entidades representativas (ou não), que seriam úteis para refinar o questionário definitivo, de modo que se pudesse melhor aferir o efeito da assimetria [Nota: a recomendação de RODRÍGUEZ (2000) era para construir as perguntas de maneira a balancear[276] as relações do tipo "é-um" e do tipo "todo-parte" entre as classes de entidades do *corpus*].

A experiência do pesquisador na dimensão espacial do problema de pesquisa e a sua vivência com o pessoal técnico que respondeu às seis perguntas formuladas nos questionários, de certa forma, ofereceram uma compensação para o relaxamento acima referido, nos seguintes aspectos:

- Nas análises qualitativas, relativamente fáceis de desenvolver, sobre algumas "surpresas" que os resultados estatísticos produziram;
- Na confirmação e ampliação das explicações, da forma mais judiciosa possível, sobre a escolha de RODRÍGUEZ (2000) pelas variáveis empíricas (subitem 5.3.2) que alimentaram as fórmulas de cálculo desse fenômeno (Equações 6.1 a 6.3) e que pareceram importantes para explicar as complexas características da SS.

Desses dois aspectos, emerge o papel de uma pesquisa exploratória como esta, ao entregar para os futuros interessados em continuá-la contribuições do seguinte tipo:

- Inferências superficiais, oriundas da análise qualitativa, que concorram para melhor explicar a natureza da SS, ainda que parcialmente (estudo-de-caso);
- Variáveis empiricamente manipuladas, que podem ser reutilizadas num estudo experimental futuro, com o intuito de verificar, mais rigorosamente, se elas são ou não diretamente responsáveis pela manifestação do fenômeno da SS.

Segundo RODRÍGUEZ (2000, p.98), trabalhos anteriores sobre a avaliação de SS, na linha metodológica de comparar o desempenho de um sistema computacional com um referencial humano de julgamento desse fenômeno, chegaram a resultados que se gruparam em duas espécies de valores de correlação. Numa escala de zero a um (domínio dos números reais), os modelos que se utilizaram do enfoque da distância semântica (puramente vetorial) não ultrapassaram um nível de correlação de 0,60. Os que utilizaram o enfoque do conteúdo informativo[277] ficaram na casa dos 0,79. Este nível subiu para a casa

[276] Mesmo sem o pré-teste, procurou-se harmonizar a escolha das classes com esse requisito.

informativo[277] ficaram na casa dos 0,79. Este nível subiu para a casa dos 0,83, quando foram utilizados modelos híbridos entre os dois anteriores.

Comparando esses três tipos de experimentos sobre SS com o seu, RODRÍGUEZ (2000) distinguiu um traço marcante do MSS[278]: o *corpus* se baseou em entidades espaciais designadas por termos de uma língua profissional (domínio de conhecimento restrito), sensível à informação contextual. Além disso, por envolver a avaliação da SS dessas entidades num quadro contextual em que os pesos da Equação 6.2 não foram arbitrados (mas determinados pelo contexto), os índices de correlação de seu modelo obtiveram um ganho substancial em relação aos modelos anteriores.

Como já explicado ao longo deste texto, para o presente estudo-de-caso, ao se atribuir pesos iguais (0,3...) para os termos da Equação 6.2, inseriu-se, de certa forma, informação contextual na avaliação da SS do PRONTO©, porque esta normalização ponderada da fórmula reflete a homogeneidade do campo humano de observações: indivíduos de mesma formação profissional, que compartilham um *corpus* homogêneo e restrito de língua profissional (*geoprocessamento*) e que até partilham um mesmo espaço de estimulação sensorial (condicionamentos militares de hierarquia, disciplina e raciocínio cartesiano), justificando-se atribuir para cada *fd* um peso igual à terça parte da unidade. Um estudo posterior, que contemplasse uma estimativa mais específica para o contexto, apoiada numa construção ontológica mais complexa, poderia concluir que uma ponderação homogênea para cada *fd* seria ou não adequada para um grupo homogêneo assim formado.

Para reforçar a conjectura do parágrafo anterior, RODRÍGUEZ (2000, p. 57-79) estabeleceu uma regra geral para a atribuição de pesos para as *fds*, na qual os pesos para cada uma das três feições é igual a 0,3..., se uma ontologia foi criada para responder a um problema específico. Se, porém, a ontologia for compartilhada, o que foge ao escopo desta pesquisa, as *fds* podem apresentar distintos níveis de relevância. Para a determinação dos pesos correspondentes a tais relevâncias, entram em jogo modelos empíricos da Teoria da Informação. Um desses modelos mede a relevância pelo grau de compartilhamento que a feição possui entre as classes de entidades do domínio de conhecimento em que se insere a ontologia. O outro modelo mede a relevância da feição em relação ao seu grau de variação (freqüência de ocorrência), no domínio de conhecimento em que se insere a ontologia.

RODRÍGUEZ (2000, p. 57-79) ainda alertou para o fato de que, tanto em ontologias específicas como nas compartilhadas, se o domínio for demasiado específico, a SS pode

---

[277] V. *modelo de conteúdo informativo* no subitem 3.2.2.2.2.

[278] E, por extensão, pode-se dizer o mesmo para o PRONTO©, exceto pela explicitação do contexto.

variar abruptamente pela simples retirada ou inserção de uma feição ou outra de uma classe de entidades, ou mesmo pela retirada ou inserção de uma classe de entidades do modelo.

Destarte, um fator determinante para atingir índices de correlação válidos em testes não-paramétricos baseia-se na correta identificação das feições distintivas das classes de entidades espaciais. Rápidas inspeções permitiram verificar que pequenas mudanças (inclusão ou omissão de *fds*) causavam grandes alterações nos resultados do PRONTO©.

Apesar de estar fora do objetivo geral estabelecido para esta pesquisa, é importante levar em consideração para pesquisas futuras que o contexto é um referencial de atribuição mais racional de pesos para os termos da Equação 6.2, contribuindo sobremodo no controle da ambigüidade criada por fatores de origem lingüística (polissemia, p.ex.), quando se tratar de compartilhamento de domínios de conhecimento.

Quanto aos índices de correlação atingidos pelo MSS, eles foram equivalentes aos do PRONTO©. No pior caso, o MSS alcançou um índice de correlação de 0,78. No melhor caso, de 0,96. Com o PRONTO©, o pior caso ficou em 0,69 e o melhor caso ficou em 0,93.

É bem provável que o menor valor de correlação de *Spearman* obtido pelo protótipo para a segunda pergunta[279] (0,69) esteja ligado à estrutura ontológica construída para a classe PÂNTANO, que seguiu o modelo conceitual da folha Faxinal. Nesse modelo conceitual, PÂNTANO é uma classe-espécie da grande categoria VEGETAÇÃO, em que pese todas as definições consultadas e a própria percepção humana de similaridade caracterizarem PÂNTANO como uma espécie de terreno (alagado), subordinada à grande categoria RELEVO. No entanto, no modelo conceitual em pauta, a classe PÂNTANO conotou brejo, lodaçal ou vegetação ligada a esse ambiente palustre.

Os indivíduos, em sua maioria, pelas respostas que forneceram, obedeceram à definição expedita fornecida em seus questionários, que também coloca PÂNTANO como "um trecho de terras inundadas", provando que, apesar de pertencerem ao grupo profissional que utiliza essa classe como uma espécie de vegetação, responderam de maneira independente, ou seguindo a definição expedita fornecida, ou seguindo a tendência generalizada de considerar a classe em tela como uma espécie de terreno (relevo/depressão).

PÂNTANO não surgiu apenas como termo-predicado (protótipo) na segunda pergunta (tendo LAGO por termo-sujeito). Ele também surgiu na primeira, na quarta e na quinta pergunta do questionário, tendo por termos-sujeitos (variantes), respectivamente, RIO, CAIS e CANAL.

---

[279] Em que o termo-sujeito ou variante foi a classe LAGO.

Por trás dos resultados numéricos apresentados nas Tabelas 6.4, 6.9, 6.19 e 6.24, que correspondem aos sumários dos casos para a primeira, segunda, quarta e quinta perguntas, nessa ordem, observou-se uma tendência do PRONTO© em pontuar a menor a similaridade de um protótipo distinto na forma (dimensões), quando comparado ao termo-sujeito correspondente, ou melhor, como PÂNTANO tem forma planar (área), o protótipo vai considerá-lo mais similar a LAGO e a CAIS do que a RIO e a CANAL; estes últimos, lineares na forma. Essa tendência do protótipo manifestou-se de forma mais ou menos regular entre os pares RIO-PÂNTANO (posto = 9,5), LAGO-PÂNTANO (posto = 5), CAIS-PÂNTANO (posto = 7) e CANAL-PÂNTANO (posto = 8,5). Como se percebe, os postos maiores (menor similaridade) acompanham os pares cuja forma é linear-planar. Evidentemente, quando um atributo se salienta na comparação entre consins (caso de "água"), essa propensão tende a ser quebrada (V. LAGO-RIO, LAGO-CANAL e LAGO-RIACHO na 2ª pergunta – Tabela 6.9).

Essa tendência, porém, não se manifestou de forma tão regular nas respostas dos indivíduos: RIO-PÂNTANO (posto = 7), LAGO-PÂNTANO (posto = 1), CAIS-PÂNTANO (posto = 9) e CANAL-PÂNTANO (posto = 7). Como se percebe no par sublinhado, ocorreu um sensível salto no critério de julgamento da similaridade entre as entidades sob avaliação, muito provavelmente em razão da presença de água nessas classes. Essa presença de água, apesar de não estar explícita na definição expedita de PÂNTANO, acompanha a representação mental que as pessoas associam a essa classe de entidades espaciais, o que pode ter produzido uma ponderação exorbitante para o seu senso de julgamento de SS.

Não foi possível estender o prazo de pesquisa para ampliar os testes do protótipo aos casos de alterações propositais na taxinomia da folha Faxinal. Se PÂNTANO figurasse noutra ordem de generalização (subespécie da grande categoria RELEVO), talvez o PRONTO© alcançasse ou superasse os índices obtidos pelo MSS. Somente trabalhos futuros nessa linha de averiguação poderão confirmar essa conjectura.

A omissão da definição do termo-sujeito ÁREA DE LAZER na terceira pergunta dos questionários distribuídos aos indivíduos não afetou os resultados de correlação em razão do perfil homogêneo dos respondentes e das próprias características das classes em comparação, todas descendendo da grande categoria INFRA-ESTRUTURA, o que lhes garantiu menores distâncias semânticas (Lei de Tobler). A confirmação disso está nas coincidências que foram mencionadas no subitem 6.4.1.3 e também na grande aproximação entre os pares formados entre ÁREA DE LAZER e os protótipos: MUSEU, REPRESA, CAMPO DE FUTEBOL, ESTÁDIO, CAMPO DE TIRO-AO-ALVO, REPRESA e VILA.

Para completar a lacuna deixada no questionário, é instrutivo expor a definição de ÁREA DE LAZER, comum aos indivíduos que compõem esse campo de observação: "Extensa área urbana, que abriga uma praça de esportes, ginásio, parque de diversão, entre outras instalações de lazer e desportivas". Apesar de não ter sido fornecida, os resultados demonstraram que, sem dúvida, os indivíduos responderam à terceira pergunta com esta definição em mente.

Surpreendentemente, apesar da omissão citada (até cogitou-se desconsiderar a terceira pergunta da análise estatística), os resultados positivos obtidos para a avaliação do termo-sujeito ÁREA DE LAZER superaram em muito os que se apresentaram para as primeira e segunda perguntas, cujas definições não faltaram aos respondentes e cuja estrutura taxinômica (exceto para PÂNTANO, como já explicado) vem de longínquo consenso no seio da comunidade *geocientífica* nacional e internacional. A explicação plausível para esse fato surpreendente, além das considerações já delineadas nos seis subitens relativos à obtenção de resultados, pode ser pautada em três itens:

- A estrutura da árvore *n-ária* (taxinomia) que reproduziu as relações semânticas entre as classes de entidades espaciais adveio de um modelo conceitual consagrado pela comunidade cartográfica mundial. Esta é uma característica nuclear do conceito de ontologia;
- Cada classe de entidades espaciais, dos níveis mais gerais aos mais específicos, foi dotada de uma quantidade de *fds* que obedeceram ao comando primordial das teorias do conceito: parcimônia informativa (generalidade, poucas *fds*) nos níveis mais altos da ontologia (maior extensão) e minúcias nos níveis mais baixos da ontologia (maior *intensão*, muitas *fds*);
- As classes de entidades especializadas da grande categoria INFRA-ESTRUTURA, à qual pertence ÁREA DE LAZER, por se tratarem de acidentes artificiais (antrópicos ou construídos pelo homem), são mais familiares aos indivíduos, particularmente aos desse grupo (engenheiros e topógrafos) do que as naturais (hidrográficas e orográficas).

Os dois primeiros itens anteriores são axiomáticos nas ontologias. Ambos parafraseiam RODRÍGUEZ (2002), que disse: "A estrutura hierárquica subjacente ao domínio da aplicação deve ser reproduzida na base de conhecimento". Pode-se até mesmo estimar numericamente esse enunciado lógico, que foi importado para a maioria das teorias do conceito. No presente estudo-de-caso, com dois campos de observação (profissionais do *Geoprocessamento* e PRONTO©), fundando-se nas categorias de alto nível de generalidade, de nível 3, das Figuras 6.10 a 6.22, 9 (nove) classes desse nível na taxinomia produziram 37 (trinta e sete) classes nos níveis mais baixos (as que foram utilizadas para o cálculo da SS situaram-se

mais entre os níveis 5 e 7). Isso quer dizer, em termos numéricos, justamente o que o enunciado anterior expressa: 20% das classes mais genéricas produziram 80% das classes mais específicas dessa base de conhecimentos.

A relação de 20%/80% para este estudo-de-caso está coerente com o experimento de RODRÍGUEZ (2002), que ficou em 30%/70%.

As considerações sobre os aparentes imprevistos ocorridos com PÂNTANO e ÁREA DE LAZER representam uma conclusão parcial de fundo cognitivo, coerente com a Lei de Tobler, em que se atribui mais semelhança às coisas mais próximas. Como as classes de entidades espaciais artificiais contribuem de forma mais efetiva para o domínio de estimulação sensorial desses indivíduos, não é mais de se surpreender que, mesmo sem definição, a avaliação de SS entre ÁREA DE LAZER com os protótipos listados na terceira pergunta alcançasse um nível maior de correlação do que, por exemplo, RIO, na primeira pergunta, com os seus protótipos.

Não fosse a implementação do módulo de edição de ontologias do PRONTO©, para montar a rede de conceitos da folha Faxinal, e talvez fosse inviável obter resultados satisfatórios para a avaliação da SS, pela inevitável fonte de introdução de erros pelo pesquisador, uma vez que a dispersão inerente ao aparelho cognitivo humano nunca seria capaz de sobrepujar as métricas algorítmicas de seqüência, condição e repetição, que foram exaustivamente exploradas para a montagem do arquivo (XML[280]) que deu corpo à ontologia *ad-hoc* da pesquisa. Esse módulo de edição foi uma das mais importantes contribuições do trabalho e o que mais consumiu tempo de desenvolvimento. Das quase dez mil linhas de código em *Java*™ (no CD – Apêndice B), pelo menos cerca de seis mil constituem esse módulo.

6.5. Resumo da metodologia

"Os números não mentem, mas os mentirosos inventam-nos."
[Gen. Charles H. Grosvenor (*apud* WONNACOTT (1980, p. 10)]

Os resultados obtidos pelo PRONTO© foram advindos da aplicação dos modelos matemáticos representados pelas Equações 6.1, 6.2 e 6.3. Tais modelos têm fundamento em conceitos de origem lógico-matemática e cognitiva, descritos no Capítulo 3 e consubstanciados numa notação formal como a BNF (V. Quadro 6.1), preenchida por relações semânticas

---

[280] V. subitem 4.3. No caso do PRONTO©, a última versão deste arquivo foi o Faxinal9a.xml.

e *fds* que os diagramas do modelo conceitual da folha Faxinal ajudaram a fornecer (Figuras 6.8 a 6.22). Esta notação, por sua vez, foi transformada num padrão de implementação (DTD), capaz de converter as sentenças lógicas da BNF para os axiomas (em número de 31) que permitiram a construção da ontologia.

Diante de todo essa sucessão de eventos metodológicos, restou apenas a confirmação estatística adequada para os resultados do protótipo. A aplicação das Equações 6.4 a 6.7 e os sumários estatísticos materializados nas Tabelas 6.11, 6.16, 6.21, 6.26, 6.31 e 6.36, constituem evidência suficiente para se afirmar que os objetivos traçados para este estudo-de-caso de manipulação experimental foram alcançados e que é possível, dentro dos limites estabelecidos neste estudo, simular o senso humano de julgamento de SS entre classes de entidades espaciais por um *software* que se enquadre nos requisitos de IA para a criação de um agente que age racionalmente (V. Figura 1.11), implementado segundo o paradigma da OO; por enquanto, o enfoque atual para LTPs que melhor exprimem aspectos do mundo por meio de objetos e de seus comportamentos (V. LRC no glossário e subitem 3.2.2.2.3).

Os requisitos básicos para a construção do segundo instrumento de verificação de hipótese desta pesquisa – o PRONTO© -, o mais complexo, não poderiam deixar de atender a todo um quadro evolutivo dos estudos do fenômeno de SS, que partiram da adaptação do conceito matemático de equivalência de RIPS (1973) e, especialmente, de TVERSKY (1977) para o domínio cognitivo, até o quadro referencial teórico atual das teorias do conceito, entre elas a do protótipo, em que se assinalam os trabalhos de Eleanor Rosch [*apud* RODRÍGUEZ, 2002)]. E em que se resume esse estado-da-arte sobre a SS? Na simples combinação da herança de *fds* entre as classes organizadas numa taxinomia. Esse esquema (na verdadeira acepção do termo – V. p.146) traduz a visão *roschiana* do conceito, segundo a qual as classes de entidades do MR são naturalmente formadas com base em seus protótipos (exemplos significativos da classe). Além do mais, as definições dessas classes devem ser focadas sobre tais protótipos (RODRÍGUEZ, 2002).

Para encerrar, fica um registro de cunho deontológico, talvez um tanto excêntrico com relação à metodologia, mas que a relativa abertura (e riscos inerentes) de um trabalho de natureza exploratória concede ao pesquisador e, também, por dever de justiça para com aqueles pesquisadores que, apesar de se arriscarem nos limites de um trabalho "politicamente correto", procuram aliar as métricas da metodologia científica com a honestidade científica, particularmente no que tange às limitações dos seus trabalhos.

Como se viu, foram impostas certas limitações aos protótipos que se pretendeu desenvolver, a fim de cumprir os cânones da Engenharia de *Software* com relação à documenta-

ção e à análise de requisitos, o que proporciona um meio mais confiável para os futuros desenvolvedores (*modeladores*) de um SIG de base ontológica, ao propiciar-lhes contornos melhor definidos do domínio da informação para o *software* que pretenderem produzir.

A construção do protótipo e mais a declaração de suas limitações são tentativas de abrir uma perspectiva mais realista na pesquisa, evitando-se o modismo nada modesto e muito arriscado de qualificar a obra como um "modelo", especialmente por aqueles pesquisadores das ciências sociais atraídos por assuntos ligados ao campo da Análise de Sistemas, mas que não buscam, ou por receio de não entender ou por vaidade, fundamentos nesse campo ou na Engenharia de *Software*, perfeitamente ao seu alcance intelectual.

Não são os esquemas gráficos esboçados no projeto, na monografia, na dissertação ou na tese, mais orientados pela subjetividade do pesquisador do que por especificações já sazonadas e consagradas pelo tempo, que imprimirão confiabilidade ao trabalho.

Onde está o modelo de um pretenso sistema ligado a *software*, se nem o preconizado protótipo foi apresentado para crítica? nas figuras? ou em estatísticas colhidas que não podem responder positivamente às seguintes questões[281], que se apresentarão na fase de testes de qualidade do *software*[282], bem depois de titulado o profissional?

- O que se quer é construir certo um produto? (verificação); ou então:
- O que se quer é construir um produto certo? (validação)

Estas são questões de ordem ética, como já mencionado, porquanto, como se falar de *software* sem se seguir essas métricas básicas?

---

[281] Colocadas por B. Boehm [*apud* PRESSMAN (1995)].

[282] Sendo otimista, se houver *software* para testar!

7. CONCLUSÃO E CONCITAÇÃO A TRABALHOS FUTUROS

"Evitar erros é um ideal pobre. Se não ousarmos atacar problemas tão difíceis que o erro seja quase inevitável, então não haverá crescimento do conhecimento."
(POPPER, 1975, p. 177)

7.1. Generalidades

O caminho pelo qual se optou para justificar a continuidade do trabalho de RODRÍGUEZ (2000) pode ter parecido um tanto desnorteante aos pesquisadores com estados de espírito avessos ou desacostumados aos meandros do pensamento filosófico, que, sem dúvida, foi a calha por onde escoaram os conhecimentos das ciências sociais e das exatas deste trabalho de natureza exploratória e interdisciplinar.

Pode-se perguntar se, para realizar um trabalho sobre informação geográfica (IG), era necessário um mergulho tão profundo em tantas obras de autores de Filosofia da Linguagem, Psicologia Cognitiva e Lingüística.

A resposta já está embutida na própria questão, já que, no Brasil e no mundo, em geral, a IG pouco foi pensada em presença de outros quadros referenciais que não fossem os das engenharias, de visão prática, objetiva e minuciosa para averiguar e dar soluções a um sem-número de problemas que gerou ao longo de mais de 30 anos de SIGs. No entanto, somente as engenharias não possuem a necessária capacidade de alcançar a verdadeira dimensão que o fenômeno geográfico irrompe e que extrapola os restritos domínios do cálculo e do apoio que presta a certos tipos de análise qualitativa.

A resposta, portanto, por tudo o que se revisou em literatura coetânea vinculada ao compromisso da convergência de métodos híbridos de pesquisa, entre qualitativos e não somente quantitativos, só poderia produzir uma resultante positiva nessa linha de pesquisa inaugurada pela Ciência da Informação Espacial ou Geográfica (CIGeo), ficando por conta e risco deste pesquisador a forma e o estilo de apresentação do conhecimento coligido e o grau de profundidade praticados.

A dose de certeza sobre excluir (ou abrandar) vulnerabilidades da metodologia adotada e incoerências do referencial teórico tem como indicador os resultados alcançados por esta manipulação experimental, orientada por uma metodologia estatística não-paramétrica que, se não é a melhor solução, é a única disponível no momento para mensuração de um fenômeno como o da similaridade semântica (SS), afastando razoavelmente bem o relaxamento do controle experimental e a negligência involuntária do pesquisador na interpretação dos

resultados alcançados, que carregaram de evidências a hipótese de pesquisa. Quanto às interpretações desses resultados que ultrapassaram as evidências descobertas, somente a continuidade deste estudo exploratório, concitada em vários trechos do texto e sintetizada a seguir, poderá demonstrar se elas foram coerentes ou não.

Já não se pode mais manter as engenharias isoladas no trato da IG. Este trabalho não é um marco nisso, nem mesmo no Brasil. O fato é que boa parte das organizações nacionais de produção cartográfica, em particular as governamentais, ainda não se deram conta dessa mudança de paradigma e, como ditam padrões de produção de *software* de *geoprocessamento*, as indústrias, também em boa parte, continuam a produzir "novas máscaras para antigas faces", remendando um módulo aqui e outro ali, no *software*, ou aumentado a sofisticação do *hardware*, acabando por construir produtos *franksteinianos*, que um dia vão inevitavelmente "se voltar contra os seus criadores".

As figuras de linguagem do parágrafo anterior não estão a criticar os métodos de produção desse *software* específico do *geoprocessamento*, porquanto seguem rigorosos padrões estabelecidos pela Engenharia de *Software*. O que se prevê é uma solução de continuidade entre a demanda de usuários cada vez mais exigentes e sem tempo para perder com aprendizado de aplicativos específicos dessas TIs e a resposta que as indústrias estão acostumadas a fornecer. Isto, sim, é o alvo da crítica.

Aprofundar, portanto, em conceitos da Filosofia da Linguagem e da Psicologia, além de recolher a julgada necessária sustentação teórica para um trabalho de um pesquisador estranho a essas áreas, foi também uma tentativa de contribuir para o estado-da-arte da CIGeo, pondo-o à disposição de usuários e produtores de IG no cenário *geocientífico* e tecnológico nacional.

Se o trabalho do qual este se originou não necessitou embrenhar-se a fundo nas áreas até há pouco tempo consideradas estranhas ao *geoprocessamento*, é porque o público-alvo da CIGeo nos EUA, Canadá e Europa já está aclimado com o estado-da-arte e com a terminologia da qual se utilizou RODRÍGUEZ (2000) para elaborá-lo. Aqui, no Brasil, e na comunidade lusófona, em geral, todavia, poucas foram as iniciativas de trabalhos nessa linha de pesquisa e, pelo que restringe o alcance desta revisão de literatura, até o momento, em português, não houve uma tentativa de levantar o estado-da-arte neste assunto.

No contexto brasileiro, o problema levantado nesta pesquisa advém de um anseio da comunidade *geocientífica* nacional que já se pode dizer antigo, quando as tecnologias da informação (TIs) começaram a ser introduzidas no ambiente de produção de documentos cartográficos, efetivamente, no início da década de 80 do século passado.

Esta evolução (revolução?) tornou possível dinamizar a obtenção de dados sobre a superfície terrestre e melhor compreender os fenômenos que a caracterizam, ampliando de maneira expressiva o conhecimento colocado à disposição do cartógrafo para representar os objetos e fatos geográficos de forma digital.

Esta avalancha de conhecimento, por outro lado, ultrapassou os limites de utilização e até mesmo de produção que eram restritos ao profissional do *geoprocessamento* e invadiu os domínios do usuário leigo em *geociências*, mercê da disseminação que é inerente às TIs, cada vez mais fáceis de manusear e menos dispendiosas nos custos de aquisição.

Foi a partir desse momento, já em meados da década de 90, que o problema da interoperabilidade entre os já maduros SIGs começaram a surgir.

O CEPAD/CONCAR (BRASIL, 1998b) foi uma resposta provisória[283] para este problema de interoperabilidade das organizações governamentais brasileiras que produzem dados e informações geográficas para a sociedade. O fulcro da questão estava na inexistência de um padrão comum de intercâmbio de dados entre as bases de dados dessas organizações produtoras.

Soluções ideais surgiram no início dos debates do foro CEPAD/CONCAR, como por exemplo, eleger uma organização responsável pela aquisição do dado cartográfico de forma única e padronizada. Sobre o acervo de dados adquiridos por esta organização, todas as outras exerceriam as suas atividades de produção.

Soluções desse tipo, muito naturais em foros de natureza científica, acabam sempre por se mostrar inexeqüíveis, quando vêm à tona aspectos de ordem tecnológica e logística.

As discussões nesse foro logo se voltaram para a manutenção das bases de dados existentes e para a concepção de um padrão comum para o intercâmbio de dados entre todas as distintas organizações. Daí, o problema passou a tomar um contorno mais definido e se apresentou sob os dois aspectos que permeiam até hoje as organizações governamentais de produção cartográfica, com os conseqüentes desdobramentos na área de pesquisa. Esses dois aspectos estão ancorados no binômio semântica – simbolização ou modelagem – tratamento gráfico-visual.

É sobre esses dois aspectos que se debruçam os partidários da Cartografia Digital (enfoque semântico) e os da Cartografia Apoiada por Computador ou CAC (enfoque gráfico-visual).

Rigorosamente, o segundo aspecto não deveria existir independentemente, visto que é parte do primeiro, a modelagem, mais abrangente.

É fato que a Cartografia Digital possui um potencial ilimitado para atender às atuais demandas por produtos cartográficos em diferentes formatos, para dispô-los em ambientes distribuídos e para orientar a busca por informações deles derivadas por meios até então inimagináveis pelos profissionais da Cartografia convencional e até da CAC, que de certa forma "congelam" a realidade em fotolitos ou arquivos digitais para os quais apenas esses profissionais tinham conhecimento sobre a sua natureza e sua produção.

Apesar de todas essas vantagens da Cartografia Digital, é prematuro vaticinar o fim da CAC, em virtude de se presumir já alcançados os propósitos para ela estabelecidos. Ainda há muita ênfase à fase de acabamento gráfico da produção cartográfica (artes gráficas), em que uma poderosa e bilionária indústria de *software* se consolidou ao longo dos anos e produziu aplicativos de automação para os tradicionais[284] processos artesanais de produção, que até hoje representam a fonte de recursos de diversas organizações.

Não se vislumbrou ainda um cenário de assunção plena e irreversível da Cartografia Digital em detrimento da CAC, até porque o enfoque desta última contribui para o avanço da primeira. Como exemplo, citam-se os vários estudos que se fazem na linha de pesquisa denominada *visualização preliminar*, conforme já se relatou neste trabalho.

Esta pesquisa, no entanto, ocupou-se de assunto ligado à Cartografia Digital, especificamente da área de maior produção científico-tecnológica - os SIGs -, em razão da natureza interdisciplinar desses sistemas.

Cabe salientar que este estudo sobre informação geográfica não implica obrigatoriamente lidar com a semântica dos elementos topológicos e morfológicos das entidades espaciais do universo geográfico. O estudo-de-caso exploratório (manipulação experimental) circunscreveu-se tão-somente aos aspectos semânticos ligados à representação terminológica (toponímia) dessa informação, coadunando o esforço de pesquisa com as contribuições que se pretende oferecer, tanto para a Ciência da Informação como para a Cartografia. Dessa forma, na linha de fronteira entre esses dois ramos do conhecimento científico, este trabalho se acomodou com razoável segurança na linha de pesquisa da CIGeo, que centros de pesquisa americanos como o NCGIA e universidades como a do estado americano do Maine mantêm em primeiro escalão de prioridade.

---

[283] Porque não se consolidou.

[284] Chamados de processos de plástico-gravura.

7.2. Tópicos relevantes da tese

Abrangência e exaustão investigativa são dois mandamentos que regem a revisão de literatura necessária para uma pesquisa exploratória. A apresentação do Capítulo 3 e do glossário (Apêndice B) representam um documento de estado-da-arte sobre a emergente CIGeo, por tratar de temas centrais para responder ao problema geral da pesquisa, numa área do conhecimento ainda em constituição, bem como por trazer entendimento mais sólido para os pesquisadores da Ciência da Informação, da Cartografia e da Ciência da Computação.

Esta tese seguiu um dos caminhos de pesquisa futura indicado por RODRÍGUEZ (2000, p. 136-141), que era o de averiguar a SS entre classes de entidades espaciais modeladas numa base de dados geográficos.

Pode-se considerar o trabalho de RODRÍGUEZ (2000) como pioneiro na investigação rigorosa da SS entre classes de entidades espaciais, ao se considerar o seu caráter híbrido, que explora tanto os modelos da Psicologia Cognitiva (modelos de feições), baseados na Teoria dos Conjuntos e em testes empíricos de estímulo-resposta, como os modelos das relações semânticas (gênero-espécie e agregação) da Ciência da Computação.

O *corpus* da pesquisa foi escolhido sem muita dificuldade, uma vez que o Exército Brasileiro, por intermédio da DSG, já havia iniciado estudos de viabilidade sobre a modelagem OO aplicada à informação geográfica desde o início da década de 90. Parte desse *corpus* foi constituída por um subconjunto do projeto-piloto referente à modelagem do espaço geográfico brasileiro: o modelo conceitual (MC) da carta topográfica da região de Faxinal (PR).

Este *corpus* da pesquisa materializou a definição de consenso para *ontologia*, no âmbito da Inteligência Artificial: "Explicitação de uma conceituação que um grupo de indivíduos partilham sobre uma parcela do mundo-real".

Para cumprir o objetivo geral delineado para esta tese, foi preciso criar um modelo semelhante[285] ao MSS de RODRÍGUEZ (2000), a fim de aferir o seu desempenho com as respostas a um questionário que foi aplicado a um grupo de engenheiros e técnicos do CCAuEx e da DSG (o referencial).

O modelo desenvolvido foi explicitado numa ontologia derivada do MC da folha Faxinal, implementado em *Java*™ (linguagem OO) e denominado de PRONTO©. Esta ontologia foi definida por um conjunto de entidades espaciais representadas por seus termos, pelas fei-

[285] Seria vantajoso utilizar o MSS, mas as dificuldades relatadas no subitem 6.1 alteraram os planos iniciais.

ções distintivas dessas classes e pelas relações lógicas de generalização e ontológicas meronímicas.

Antes do PRONTO©, outro modelo – o PROFAX - foi investigado, baseado apenas no enfoque vetorial.

O PROFAX, também implementado em *Java*™, mostrou-se incapaz de avaliar semanticamente classes de entidades espaciais representadas por termos individualizados e independentes. Ele só se mostrou capaz de avaliar a SS entre conjuntos ("pacotes") de classes representadas pelos respectivos conjuntos de termos. Mesmo assim, este protótipo revelou o potencial desse mecanismo de determinação de SS e preparou o caminho para implementar um modelo mais aprimorado, implementado pelo PRONTO©, que combinou o enfoque do seu predecessor (distância semântica numa taxinomia) com a modelagem matemática de RODRÍGUEZ (2000, p. 53-57), aplicada às feições distintivas de classes de entidades espaciais denotadas por termos (com base na Teoria dos Conjuntos) e às relações entre essas classes.

Os fundamentos desses enfoques foram extraídos pela IA de várias disciplinas de natureza cognitiva, nomeadamente da Ciência da Computação e da Lingüística Computacional, sendo oportuno assinalar as principais características dessa ferramenta desenvolvida, orientada para estender o MSS:

- Formalização matemática de parte dos processos cognitivos humanos ligados ao julgamento de similaridade semântica entre classes de entidades espaciais, ou melhor, da maneira pela qual as pessoas se comunicam em relação aos conceitos de natureza espacial, num domínio restrito do conhecimento;
- Pelo enfoque desta tese, a avaliação de SS é um processo em que são analisadas as propriedades e características (comuns e distintas) de classes de entidades espaciais, bem como as relações entre essas entidades;
- O desenvolvimento do protótipo contemplou parcialmente os três níveis de interoperabilidade para SIs, na ordem crescente de conteúdo informativo: o ***sintático*** (estruturas de dados clássicas para carregar as definições); o ***esquemático*** (definições das classes e a explicitação das suas organizações hierárquica e de agregação); e o ***semântico*** (definições das relações aplicadas às instâncias das classes para reproduzir o comportamento das entidades o mais próximo possível do mundo-real).
- Incorporação da seguinte definição operacional de ***ontologia***: "Estrutura que captura a visão do mundo, que permite buscas *intensionais* do conteúdo informativo de uma BD,

que define a semântica dos dados, independentemente de suas relações sintáticas e esquemáticas e que reproduz a relevância dos dados sem necessitar de acesso a eles".

Com relação à primeira característica levantada anteriormente, cabe relembrar um ponto ligado à parte cognitiva do modelo, que repassou subsídios à parte matemática do PRONTO©. Os modelos matemáticos de RODRÍGUEZ (2000) refletem os fenômenos cognitivos da assimetria e da distância semântica que estão subjacentes à estrutura hierárquica adotada. Mas isto não se fez de forma arbitrária e sim de forma empírica, por meio de testes psicológicos baseados no tempo de reação (resposta) a certos estímulos de associação entre dois conceitos. Quanto mais rápido for o indivíduo na sua resposta, menor será a distância semântica na representação mental que o indivíduo estabelece entre o termo variante (sujeito) e o protótipo (predicado). Se muitos indivíduos respondem da mesma forma, forma-se um padrão, que está ligado à *tipicidade* do conceito (Teoria do Protótipo).

7.3. Resultados alcançados

Os resultados obtidos no Capítulo 6 desta tese permitiram responder ao problema geral de pesquisa formulado no Capítulo 2 e corroborar a hipótese alternativa do Capítulo 5, ou seja, é possível obter resultados plausíveis por um *software* que simule o senso humano de julgamento de similaridade semântica de entidades espaciais.

Da mesma forma acima, os resultados obtidos pelo PRONTO© e a sua aferição por métodos estatísticos não-paramétricos, configurados para níveis de significância toleráveis para o tipo de manipulação experimental que se levou a efeito, permitem asseverar que os objetivos específicos estabelecidos no subitem 5.2 foram atingidos, o que, conseqüentemente, corroborou as hipóteses estatísticas do subitem 5.3 e, fechando o ciclo metodológico da pesquisa, garantiu a consecução do seu objetivo geral. Destarte, ficou evidenciada a coerência ou plausibilidade da metodologia adotada nesta pesquisa.

Como a presente pesquisa é de natureza exploratória, é de se esperar uma numerosa lista de proposições para trabalhos futuros, o que virá no subitem 7.5. Além disso, por ter enveredado pelo universo cognitivo da linguagem, a metodologia foi também afetada por análises qualitativas, que as conclusões a seguir ensejam pôr em relevo.

Em síntese, o presente esforço se junta aos congêneres de outros pesquisadores, aqui citados ou não, que intentam contribuir na concepção de modelos e no projeto de sistemas de informação em que o conceito e a aplicação da similaridade semântica desempenhem um papel fulcral para aproximar cada vez mais usuários ávidos por informação geográfica,

sejam esses usuários profissionais das ciências exatas e das engenharias, profissionais das ciências sociais ou mesmo leigos nesses dois grandes ramos das ciências, simplesmente curiosos nas tecnologias de *geoprocessamento*.

7.4. Conclusões

O preâmbulo deste subitem apresenta uma conclusão extensa o suficiente para afetar todas as outras que se seguem.

Esta conclusão abrangente relaciona-se à determinação de parâmetros como o da SS entre termos que denotam classes de entidades espaciais numa ontologia, a chave para aprimorar os SIGs nos dois níveis mencionados por RODRÍGUEZ (2000), quando se analisou a tendência desses sistemas para a próxima década, quais sejam: o nível da recuperação de informações e o da integração de informações.

No primeiro nível citado, o interesse está em prover meios de buscar (descobrir) informações em acervos de dados da forma mais intuitiva possível para os usuários. No segundo caso, o interesse está na identificação de objetos semanticamente similares em diferentes BDs e, assim, comparar os modelos de dados dessas bases, integrando-as e permitindo o intercâmbio de dados entre sistemas de diferentes configurações.

No campo das realizações práticas no ambiente de desenvolvimento e produção de SRIs, KIRYAKOV (1998) e CRANEFIELD (2002) dão a medida exata do que se enquadra nesses dois níveis.

Bem típico do primeiro nível, o trabalho de KIRYAKOV (1998) distinguiu nesses sistemas módulos de reconhecimento de elementos textuais de extração de significado (EATs: entidades atômicas de texto), utilizados para calcular a semelhança entre o texto da pesquisa com o armazenado num repositório de dados de alta estruturação (base de conhecimentos). Esta concepção foi revisada nos trabalhos de diversos autores [(JIANG, 1997), (RESNICK, 1999), (WONG, 2000), (SANTOS, 2002) e (GANESAN, 2002)], tratados no subitem 3.2.2.2.2, cujos fundamentos foram aplicados no PROFAX.

A tendência dos modelos concebidos para explorar esse nível de recuperação de informações está em embutir conceitos nas EATs, de forma que elas sejam capazes de capturar expectativas subjacentes e de valor intuitivo dos usuários e que não estão objetivamente presentes numa consulta. É a decantada "descoberta de informação" de RODRÍGUEZ (2000), meta que cada vez mais vem substituindo o objetivo quase esvaziado de sentido de "localizar dados", assunto examinado no subitem 2.2.3 por BORGES (2002).

CRANEFIELD (2002), já no nível de integração de informações, ocupou-se de ferramentas ou aplicativos de especificação de ontologias, exaltando as vantagens que as tecnologias OO oferecem na forma de linguagem de representação de ontologias (LRO).

Os trabalhos de CRANEFIELD (2002) e de SILVA (2001) alinham-se com a tendência de sinergia entre as técnicas de modelagem OO e as técnicas de classificação, particularmente fundadas na Terminologia (V. Quadro 3.4), revelando que a arrazoada crítica de SÖRGEL (1999) talvez passe a não mais ter sentido nas pesquisas interdisciplinares da Ciência da Computação e da Ciência da Informação, quando o problema se situar no tema da SS, envolvendo aspectos de recuperação e integração de informações.

Essa tendência à perseguida sinergia de SÖRGEL (1999) parece ter germinado um novo campo de pesquisas e de aplicações. Vários autores no subitem 3.2.2.2.3 concluíram sobre o papel das ontologias e das LROs como elementos indutores de uma área de engenharia emergente: a ***Engenharia de Ontologias*** ou Engenharia Ontológica, cujo objeto, ainda que nebuloso em seus primórdios, parece se delinear na forma de um conjunto de técnicas e de conceitos colocados à disposição de profissionais e pesquisadores interessados no desenvolvimento de linguagens destinadas à modelagem rápida e sistematizada de ontologias, assim como na utilização dessas linguagens para participar de um novo ciclo de vida de sistemas de base ontológica, particularmente em razão dos seguintes fatores positivos:

- ***Comunicação:*** ontologias são ferramentas úteis para ajudar as pessoas a se comunicarem, sob várias formas, acerca de um determinado assunto (domínio do saber). Talvez por isso, segundo DIAS (2003), a Engenharia Ontológica seja mais acessível aos pesquisadores das ciências sociais aplicadas do que aos da Ciência da Computação. Estes últimos procuram entender esses novos conceitos por um prisma um tanto reducionista, tentando explicitá-los, imediatamente, no âmbito da OO, por meio de gráficos ou diagramas que expliquem as propriedades estáticas das entidades que representam e o dinamismo de seus comportamentos, ou mesmo por um DER (diagrama de entidade e relacionamento) ou DFD (diagrama de fluxo de dados), no âmbito da análise estruturada;
- ***Formalização:*** em razão da natureza formal da notação utilizada, a especificação de um domínio do conhecimento elimina as contradições e inconsistências que o efeito da ambigüidade normalmente produz numa especificação de características descritivas (em língua natural). A especificação formalizada é o primeiro passo para a carga de uma ontologia num sistema computacional (automação), o que propicia uma sólida metodologia científica de verificação e validação da representatividade da ontologia criada;

- ***Representação do conhecimento e reutilização:*** uma ontologia encerra em sua estrutura um vocabulário de consenso e representa o conhecimento de um domínio do saber de forma explícita e no seu nível mais alto de abstração, tornando-a simples de entendimento, além de contar com uma enorme capacidade de reutilização. Esse fator é imprescindível para a almejada interoperabilidade dos SIs do século XXI.

As conclusões que se seguem obedecem ao critério de seleção das conclusões parciais que surgiram ao longo do texto, particularmente as do subitem 1.5.4, do Capítulo 3 e do Capítulo 6, que se coadunam com os dois níveis funcionais anteriormente citados e estabelecidos para SIs. Muitas dessas conclusões são extensões das de outros autores e constituíram embriões para alguns dos trabalhos futuros propostos no próximo subitem.

Assim como concluiu MEDEIROS (1999, p. 253-259), esta pesquisa também pode ficar sujeita à crítica mais comum nesse assunto: a da restrição a determinado domínio de conhecimento em que se fixou. No entanto, como já alertaram outros autores da IA (RUSSELL, 1995), não há como se evitar esta limitação. Por enquanto, só se pode simular por IA ações racionais ou *agentes que se comportam com racionalidade* (V. Figura 1.11) e aplicações dessa espécie só podem ocorrer em *corpora* muito limitados do saber.

Sem sombra de dúvida, é mais promissor em termos científicos e, particularmente, em termos tecnológicos, lidar com fenômenos que possam ser formalizados e mensurados no campo do possível, do que ficar no ambiente ideal das entidades que habitam somente o campo da imaginação.

A idéia delineada nos dois últimos parágrafos é bem coerente com a realidade, uma vez que a IA tem oferecido produtos de benefício tangível para a sociedade.

Apesar da natureza dinâmica dos conteúdos das bases de conhecimentos que se formam, graças a esses sistemas de IA que exploram a semântica implícita nos dados, suas formas de organização obedecem a certos padrões que podem definir estruturas universais de intercâmbio. A SS incorporada numa rede semântica, a semelhança do Zstation™ do Prof. Henri Zinglé [*apud* MEDEIROS (1999)], do MSS de RODRÍGUEZ (2000) e do PRONTO©, nesta pesquisa, é um mecanismo muito promissor para revelar tais estruturas de intercâmbio.

Essas últimas conclusões parciais de autores que produziram resultados práticos com SIs apoiados na IA parecem ir ao encontro da posição de SCHANK (1995), que criticou tanto o lingüista Noam Chomsky como o filósofo John Searle em relação às suas posições contrárias à hipótese da funcionalidade (V. subitem 1.5.4), tão explorada pela IA e fonte da maior contribuição desta disciplina para os sistemas de informação, qual seja: a criação de *mode-*

los de processos das *aptidões mentais*. Tem sido provado que esses modelos (tanto na IA como em campos aliados) podem ser simulados em sistemas computacionais.

Conclui-se, outrossim, que a metodologia adotada neste trabalho pode ser aplicada noutras áreas do conhecimento. Há, evidentemente, necessidade de coletar e organizar conceitos para a construção de ontologias específicas (*ad-hoc*), segundo o domínio de estudo. Essa prática, segundo MEDEIROS (1999, p. 253-259), é comum na Ciência da Informação, em prol da precisão no nível de recuperação de informações.

As conclusões anteriores acabam por consagrar as recomendações expostas no artigo de BÄHR (1996), que podem ser sintetizadas na sua preocupação em não se poder tratar do conteúdo informativo de imagens, fotos digitais e outros produtos afins somente pelas *geometrias*. É curiosa e reveladora esta conclusão, pela coincidência com o objetivo estabelecido nas reuniões do grupo CEPAD/CONCAR (BRASIL, 1998b), já que é muito improvável que tivesse ocorrido conhecimento sobre o artigo do autor, em 1996, por algum dos componentes daquele grupo no início dos trabalhos, que ocorreram a partir de 1996. Isto revela a irrefreável tendência do pensamento científico atual para estabelecer a interdisciplinaridade entre o *Geoprocessamento* e as ciências cognitivas.

Há uma relação íntima entre a linguagem e o mundo-real (V. subitem 3.1.3). A questão que se colocava é a de como encontrar uma teoria lingüística que fosse adequada à recuperação da informação contida nesses documentos modernos e como esta teoria poderia dar suporte à interpretação desta informação recuperada, criando uma teia de conceitos na cadeia de produção de conhecimento. Nesse ponto, não se pode mais depender apenas das técnicas de *geoprocessamento*. É hora de se socorrer nas ciências cognitivas.

Sendo assim, conclui-se que é fundamental distinguir entre as capacidades da atual geração de SIGs e as limitações inerentes a qualquer representação computacional do espaço geográfico. Em que pese ser inexeqüível capturar num ambiente de *geoprocessamento* todas as dimensões de conceitos como *sistemas de objetos* e *sistemas de ações*, é importante buscar técnicas que propiciem aproximações dessas dimensões. Para tanto, será necessário utilizar enfoques quantitativos baseados em conceitos sobre *ontologias* e *representação do conhecimento*, sem perder de vista que os modelos oriundos desses enfoques serão, no máximo, clones da realidade geográfica. Esta posição é compartilhada por alguns cientistas brasileiros[286] que pesquisam SIGs.

---

[286] Citados na trilogia de CÂMARA (2002).

Uma outra conclusão que se tira neste trabalho liga-se à importância do *contexto* na avaliação da SS e que vai gerar recomendações a pesquisas futuras no próximo subitem.

No subitem 3.2.2.1, BÄHR (2000) resgatou a importância das *relações* em quatro tipos de ferramentas de representação do conhecimento, entre elas a *rede semântica* (RS), explorada no PRONTO©. O autor salientou que, em geral, essas ferramentas concentram-se mais nos nós das redes (conceitos) do que nos arcos que ligam esses nós (as relações). Dessas comparações, BÄHR (2000) concluiu que o fenômeno espacial, assim como o verbal, é muito dependente do contexto, que pode ser adequadamente tratado pelos arcos das RSs.

O papel do contexto na solução de ambigüidades é reconhecido na literatura por diferentes autores [MEDEIROS (1999) e RODRÍGUEZ (2000)]. Uma forma ambígua (termo ou texto) pode ser interpretada de maneira unívoca quando utilizada em determinado contexto. Em sistemas de tratamento da linguagem natural, a ambigüidade não tratada pode provocar um encadeamento de análises incorretas.

As duas autoras citadas no parágrafo anterior tentaram modelar uma forma de controle do efeito nocivo da ambigüidade na precisão da recuperação da informação, dentro dos objetivos específicos que estabeleceram para os seus estudos; MEDEIROS (1999), com foco em elementos frasais, utilizou o sistema Zstation™; RODRÍGUEZ (2000), com foco em elementos terminológicos, desenvolveu o MSS. Esta última autora relatou que houve um ganho substancial nos resultados de SS em relação aos trabalhos antecedentes, quando foi carregada informação contextual em sua ontologia.

Neste estudo-de-caso, como já explicado, foram atribuídos pesos iguais (w_p, = w_f = w_a = 0,3...) para os termos da Equação 6.2. Esta atribuição não deixou de ser uma forma de inserção de informação contextual na avaliação da SS do presente modelo, justificada pela homogeneidade do universo lingüístico e até comportamental dos indivíduos escolhidos para os testes.

Quanto aos índices de correlação, os atingidos pelo PRONTO© em nada ficaram a dever ao MSS de RODRÍGUEZ (2000). O primeiro protótipo alcançou uma média de correlação de 0,81 e o segundo, uma média de 0,87. A superação do segundo em relação ao primeiro deveu-se à introdução de axiomas na ontologia que modelaram os pesos segundo o *enfoque contextual da substituição* (MILLER, 1981). Mas é preciso acrescentar uma nota sobre esta aparente grande vantagem do modelo conceitual que utilizou o PRONTO©. O fato é que, como diversas vezes já mencionado no texto, a capacidade de avaliação de SS deste modelo foi majorada em virtude de alguns fatores. Desses fatores, um teve papel relevante no resultado. Pode-se inferir que a seleção do grupo de indivíduos (campo de observação)

selecionados para o estudo-de-caso, muito homogêneo, adequou a realidade ao modelo matemático, ou seja, se ajustou ao critério arbitrário de distribuição equânime dos pesos atribuídos aos três coeficientes de peso da Equação 6.2. Tais pesos representam justamente parâmetros de entrada para a informação contextual no modelo. No próximo subitem serão propostos dois trabalhos pertinentes ao assunto.

Do que se tratou no subitem 3.2.2.2.1 (p. 134), em que foram revisados assuntos de Filosofia da Linguagem, ficou bem claro que o significado é imanente à expressão lingüística, independente do mundo empírico, e que possui um valor dentro do sistema de LN à qual pertence a expressão (termo). Donde se conclui que, se este valor está inserido num sistema, é possível medi-lo, ou mais precisamente: avaliá-lo relativamente a outros valores (nível da *conotação*), como a presente tese teve por objetivo alcançar.

PINTO (1977), forneceu subsídios para a construção de uma teoria semântica que produzisse resultados tangíveis (V.subitem 3.2.2.2.2). Ele via os *significados* como classes de entidades portadoras de informação, interpessoais e sociais, que garantem o funcionamento da língua como instrumento de comunicação. Esta definição integra várias visões de filósofos da linguagem e semanticistas. O autor, adepto convicto da linha de pesquisa caracterizada pela "quantificação" do significado, preconizou princípios gerais para a sua avaliação (relativa), chamando-a de *cálculo de intensões* ou de *predicados*, por ser a *intensão* ou conteúdo informativo a fonte mais viável e plausível de investigação para a Lógica.

PINTO (1977) só não imaginava, àquela época, que resultados bem-sucedidos só viriam a surgir quase trinta anos depois e que trabalhos no campo da Cartografia somariam esforços de pesquisa para a sua decantada teoria. A comprovação disso está na síntese da revisão de literatura, em que o ponto comum do material revisado das *geociências*, Ciência da Computação (IA) e outras ciências cognitivas é a tendência de convergência das pesquisas. Uma terminologia comum está paulatinamente sendo formada.

Outra conclusão que se tirou no Capítulo 3 (subitem 3.2.2.2.1, p. 134), que aduziu para o plano de revisão de literatura a atribuição de estender o estudo para áreas aparentemente estranhas ao objetivo geral desta pesquisa, foi baseada na análise lógica (*modus ponens*) do argumento de PINTO (1977) e de RODRÍGUEZ (2000), também acolhido nesta tese, de que um termo é denotativo para uma classe de entidades espaciais, se ele possui um conjunto suficiente de feições distintivas que caracterizem a classe da qual é referente, no mesmo nível de abstração de outras que lhe são assemelhadas ou vizinhas conceitualmente.

Sendo assim, o argumento que rege esta pesquisa é logicamente plausível e, no aspecto científico, este argumento necessita de apoio metodológico para comprovar a hipótese

de pesquisa. A metodologia revelou a verdade das premissas, i.e., a metodologia desta pesquisa demonstrou a existência de evidência nas premissas que se formularam: as definições das classes de entidades espaciais foram suficientemente adequadas (claras e distintas).

Em reforço a uma parte da conclusão anterior, vem a recomendação de POPPER (1975, p. 170) para empreender uma investigação profunda e até histórica a respeito de todos os aspectos ligados a uma teoria e ao seu problema, incorporando este esforço ao objetivo da pesquisa, o qual se manifestou de forma tácita no presente trabalho.

Foi dessa revisão de literatura diversificada que se concluiu que a SS não goza de todas as propriedades de uma relação matemática de equivalência[287]. Há aspectos de ordem cognitiva que entram em jogo, conforme as conclusões de diversos pesquisadores sobre o fenômeno da SS, entre os quais se distinguem Eleanor Rosch [*apud* RODRÍGUEZ (2000)], TVERSKY (1977), KRUMHANSL (1978), RADA (1989) e SONESSON (2002).

Sintetizando as pesquisas de todos esses autores, pode-se dizer que a SS equivale a um critério de ordenação de forte natureza cognitiva e, por isso, irredutível, em sua totalidade, a fórmulas matemáticas.

Experiências de psicólogos para aferir os tempos de resposta de indivíduos, em exercícios de comparação entre pares de objetos ou seres, estabeleceram os parâmetros que adaptaram as formulações determinísticas da equivalência matemática e da distância semântica num espaço vetorial à natureza cognitiva da similaridade semântica.

Desses experimentos, surgiram primordialmente três linhas de pesquisa na tentativa de modelar matematicamente a SS: a que se associou à *Teoria dos Conjuntos*[288], em que as unidades de trabalho sujeitas às funções dessa teoria (interseção, união, diferença) são as *feições distintivas*; a que se associou à *Geometria Analítica*[289], em que os termos podem ser representados por vetores num espaço euclidiano e organizados noutras estruturas de representação do conhecimento (rede semântica - RS); e, finalmente, a *linha híbrida*, adotada neste trabalho, que se apropriou de conceitos e técnicas das outras duas linhas e que pode ser explicitada no seguinte enunciado: *"Quanto mais propriedades em comum possuírem dois objetos, mais ligações numa RS eles terão e mais similares também serão"*.

Portanto, não se deve confundir a relação de ordem cognitiva da SS com a formal de equivalência, desde que o signo lingüístico é assimétrico e não-reflexivo, não se podendo defini-lo em termos essencialmente lógicos. Confundir similaridade com equivalência num

---

[287] Relações de equivalência: Igualdade, similaridade, identidade e diferença.
[288] TVERSKY (1977).
[289] De RADA (1989) e de sua equipe.

contexto de rigor científico, seria admitir que o homem vive no mundo abstrato das ciências exatas. No entanto, o homem vive num meio sócio-cultural bem mais complexo que o anterior, em que a SS é um fenômeno de ordem cognitiva, confirmado por testes laboratoriais de estímulo-reposta e que só pode ser formalizado em casos ainda muito restritos.

As conclusões seguintes referem-se à síntese das conclusões parciais tiradas no subitem 6.4.2, algumas das quais já foram comentadas *en passant*, com ênfase no contexto. Desses comentários conclusivos anteriores, pode-se juntar a ilação de (MILLER, 1991, p.3), confirmada pela Equação 6.2, de que *"A SS é uma variável dependente do contexto"*. Para este autor, *a representação contextual de um termo*, reduzida aos três coeficientes de peso da Equação 6.2, *constitui o conhecimento de como este termo é empregado*.

7.4.1. Conclusões sobre a extensão do modelo de avaliação de similaridade semântica

O desenvolvimento de um *software* como o PRONTO©, a criação de uma ontologia específica para o estudo-de-caso exploratório em questão e ainda a aplicação de um questionário para obtenção de um referencial humano de comparação com o instrumento (meio) automatizado de pesquisa só assumem o seu papel de coadjutores no processo de avançar as fronteiras do conhecimento nessa área interdisciplinar de interesse para a Ciência da Informação, em que operam as ciências cognitivas e as geociências, quando se torna possível tirar conclusões de caráter restrito, decorrentes da análise dos resultados que foi realizada no subitem 6.4.2. Por conseguinte, coroando este esforço de pesquisa, estas conclusões foram salientadas neste subitem, algumas das quais já poderão ser postas em prática no ambiente de desenvolvimento de modelagem do espaço geográfico brasileiro[290], assim como em ambientes acadêmicos de pesquisa, pelo interesse de alguns pesquisadores na área de ontologias. As conclusões de caráter restrito vêm relacionadas a seguir:

- A Toponímia foi inadequadamente instituída como categoria no MC da folha Faxinal. Nessa fase da gênese de um SIG, em que o foco da análise deve se voltar para a visão de um mundo que vai muito além do mero *grafismo* (simbolização final dos fenômenos geográficos), a "visão" desse MC deve orientar-se para as entidades espaciais genuínas e para a descoberta das ações e reações a que estão sujeitas;
- O padrão de definição de tipo de documento (DTD), implementado num arquivo XML™, foi um mecanismo eficiente de edição da ontologia *ad-hoc* desta tese, combinando a construção da arquitetura taxinômica da folha Faxinal com a inserção de trinta e um axi-

---

[290] A cargo da 1ª DL (RS).

omas de natureza sintática e semântica (*estruturas formais da definição predicativa de cada classe*);

- O projeto e a implementação do módulo de edição da estrutura ontológica do PRONTO© coaduna-se com os esforços do OMG pela busca de um padrão de intercâmbio de modelos de geração de enunciados em LRO (o padrão XMI™);
- Em função do conteúdo do item anterior, o módulo de edição do PRONTO© foi projetado para utilizar um arquivo que armazena uma DTD, no formato XML™, compatível com a maioria das LTPs de domínio público (no caso presente, o *Java*™). Esta DTD contém a hierarquia da ontologia (relações "é-um" e "todo-parte"), assim como os trinta e um axiomas das *estruturas formais de definição predicativa* (*fds* contidas nos *consins*) de cada classe de entidades espaciais;
- Pequenas mudanças nas *fds* (retirada ou inserção) das classes de entidades, produzem efeitos sensíveis no cálculo da SS em protótipos como os que implementaram o MSS e o modelo de extensão desta tese, especialmente se o domínio de conhecimento em que se insere a ontologia for restrito;
- Se o modelo conceitual que constituirá parte do *corpus* da ontologia não expressar o modelo mental humano de processamento de informação (percepção, indução de propriedades e abstração na conclusão), provavelmente o cálculo da SS para as classes de entidades envolvidas será afetado por erros de natureza imprevisível (V. subitem 6.4.2 – comentários sobre a classe PÂNTANO);
- Uma ontologia construída com base no princípio da parcimônia informativa para as classes mais genéricas e na pormenorização informativa crescente, em direção às classes de menor nível na taxinomia que a compõe, significa estabelecer uma relação positiva entre o modelo de especificação de um domínio de conhecimento representado pela ontologia, e a "visão" dessa porção da realidade, partilhada por um grupo de indivíduos.

Assim como MEDEIROS (1999) e RODRÍGUEZ (2000) concluíram nos estritos domínios de seus estudos-de-caso, também concluiu-se nesta pesquisa que a estrutura taxinômica[291] escolhida para representar os conceitos das entidades espaciais envolvidas no estudo, bem como a definição adequada (clara e distinta) dos termos (toponímia) constituem fatores determinantes para o resultado geral obtido pelo PRONTO©, traduzido na corroboração da hipótese alternativa de pesquisa, i.e., de que é possível solucionar por meio de procedimentos automáticos casos de similaridade semântica entre classes de entidades espaciais, mo-

[291] Uma arquitetura de árvore *n-ária* para reproduzir os conceitos (e suas relações) de uma rede semântica.

deladas conceitualmente segundo o enfoque da OO e representáveis em bases de dados cartográficos digitais.

7.5. Recomendações para estudo (trabalhos futuros)

Neste subitem, as recomendações para o prosseguimento da pesquisa foram grupadas em dois ramos: o primeiro voltado para trabalhos de natureza teórico-exploratória e o segundo, para os de natureza experimental.

7.5.1. Trabalhos teórico-exploratórios

Nesse caso, recomendam-se dois trabalhos: um de síntese ou de estado-da-arte sobre a evolução da SS no domínio das ciências cognitivas[292] e um segundo que trate de outro tipo de lógica para formalização de problemas complexos que envolvam sistemas de informações e aspectos cognitivos de produtores e usuários de informação geográfica.

7.5.1.1. Prolegômenos sobre o estado-da-arte da CIGeo

O objetivo desse estudo proposto deve pautar-se por aperfeiçoar ou até mesmo redirecionar o arcabouço conceitual sobre o qual se assentam as pesquisas da CIGeo, ampliando o alcance de uma das primeiras iniciativas nesse esforço em língua portuguesa, consubstanciada na trilogia de CÂMARA (2002), assim como ampliando o ensaio que aqui se fez no subitem 1.5.4 e no Capítulo 3, com reforços eventuais do glossário (CD – Apêndice B).

Além disso, o trabalho deve contribuir com novos enfoques para os problemas de recuperação e integração de informação geográfica (ou de outra espécie), tendo em mira formulações mais simples e elegantes para a avaliação da SS, ou que melhor descrevam o *contexto* - variável independente da SS -, de tal sorte que os modelos porventura derivados desse estudo possam ser implementados, no futuro, a baixos custos em matéria de esforço computacional (tempo de execução e memória).

Se os trabalhos mais recentes sobre o tema explorado nesta tese colocam um fim às críticas de SÖRGEL (1999) sobre a falta de sinergia entre as ciências da classificação (Ciência da Informação) e as ciências exatas (Ciência da Computação), por outro lado MILLER (1991, p.3) já reclamou da falta de uma revisão de literatura extensiva sobre a SS e o papel do contexto que contribua para o acerto dos rumos das ciências cognitivas (em particular, da

[292] Na acepção de ABRANTES (1994, p. 9): Lingüística, Ciência da Computação, Psicologia Cognitiva e Neurofisiologia.

Psicologia Cognitiva) e, pelo que se depreendeu da revisão de literatura realizada nesta pesquisa, o quadro não se alterou muito desde os primórdios da década de 90 (séc. XX).

Ao longo de toda a revisão de literatura (Capítulo 3), especialmente no subitem 3.2.2.2.1, uma preocupação esteve sempre presente: o grande desafio para criar (esboçar) uma teoria semântica, porque se trata de conciliar aspectos até hoje aparentemente inconciliáveis entre os objetos das ciências sociais, de onde se originaram os problemas, e os modelos da Lógica e da Matemática, às vezes inconsistentes com a realidade complexa que os pesquisadores procuram sistematizar.

No entanto, trabalhos na linha de pesquisa da SS entre classes de entidades espaciais contribuem objetivamente para a formalização de uma teoria semântica.

As esparsas reflexões, considerações e esboços de modelos de semanticistas como PINTO (1977), assim como os resultados de estudos parcialmente quantitativos nessa área, como o de RODRÍGUEZ (2000) e esta própria pesquisa, com todas as limitações, são indicações plausíveis de que já há material teórico e empírico para ser consolidado num estudo de síntese que, sem dúvida, constituirá o ponto de partida para a concepção da teoria semântica reclamada por PINTO (1977) e que muito contribuirá para todas as áreas que com a Lingüística mantenham interação.

7.5.1.2. Estado-da-arte sobre a similaridade semântica de entidades espaciais no contexto de uma teoria semântica

Recomenda-se que a continuação de estudos nessa linha proposta avalie as concepções sobre o *significado* e suas características intrínsecas de *intensão* e denotação num sistema de conceitos, nos aspectos atinentes ao desenvolvimentos de SIs, em geral, e nos SIGs, em particular.

A conclusão de PINTO (1977), no Capítulo 3, indica que ainda não é possível formular uma teoria capaz de predizer o sentido de uma sentença isolada com base nos itens lexicais presentes e nas relações sintáticas entre eles, com base em termos ou semas de uma LP, mesmo na presença de contexto ou, o que é ainda mais complexo, com base em texto-livre de assunto geral (frases em LN).

Em razão disso, os pesquisadores da área lingüística poderiam ater-se a uma avaliação circunstanciada do ideal *transformacionista* de precisão nas avaliações de significado, baseadas na tese dos chamados "universais semânticos", aplicados em *corpora* de LN. PINTO (1977, p. 25-90) mostrou-se céptico com relação a esta tese e já advertia que até mesmo as avaliações se significado de natureza genuinamente semântica são estimativas

(aproximadas). Uma pesquisa que confrontasse essas duas posições, fortalecendo uma delas ou estabelecendo um novo enfoque conciliador, sem dúvida alguma, contribuiria sobremaneira para o avanço de uma teoria esclarecedora sobre a natureza do significado.

Caso o pesquisador da área da Lingüística também seja familiarizado ou mesmo especializado na Ciência da Computação (Lingüística Computacional), a concitação recomendada *ut retro* ganhará uma nova dimensão, mais rica em recursos de mensuração e de subsídios para uma análise mais profunda do significado, como vêm demonstrando os resultados obtidos de sistemas que exploram apenas termos de *corpora* mais específicos (LP), como nesta pesquisa e na de RODRÍGUEZ (2000), ou mesmo de sistemas que exploram *corpora* compostos de frases completas, como no trabalho de MEDEIROS (1999).

O que se espera de pesquisas sobre a natureza do significado, com o auxílio da computação, é que surja uma teoria ou pelo menos um conjunto de hipóteses plausíveis que se coadunem com o ideal primordial perseguido por Gottlob Frege: "Revelar a verdadeira estrutura lógica que está por trás das sentenças da língua natural, que às vezes é muito diversa da sua estrutura aparente".

O que o pesquisador que optar por esse tema proposto não deve perder de vista é que um trabalho de estado-da-arte, nesse nível, não deve ater-se apenas à descrição da evolução dos campos do conhecimento que confluíram na SS. É preciso ter cuidado com os aspectos da pesquisa que confinarem com a Filosofia, especialmente a Filosofia da Linguagem e a Filosofia das Ciências. Como o estudo é de caráter transdisciplinar (V. glossário), obviamente a incursão do pesquisador em fronteiras do conhecimento que não lhe sejam familiares, nem sempre estará isenta de imperícia e, nas palavras do filósofo Jean Rostand [*apud* HUISMAN (1976)], até mesmo de certa ingenuidade. Cabe, então, ao pesquisador, recorrer à Filosofia e à Epistemologia, a fim de retomar o rumo sóbrio e objetivo de seu estudo e para delimitar o que se poderá concluir no estrito alcance de sua autoridade.

7.5.1.3. Estudo de outro gênero de lógica para mediar o mundo factual no processo de explicitação de uma ontologia

RUSSELL (1995), em que pese identificar diversas limitações da LPO para representar a realidade empírica, admitiu que este é o tipo de lógica ainda mais difundido e utilizado na formalização de AIs[293].

[293] V. posições *cognitivista* e *conexionista* (1.2.2) e noção de modelo lógico (1.5.3.1).

O que se propõe como pesquisa de cunho exploratório é estudar a possibilidade de emprego de um outro modelo para formalizar ontologias, baseado noutras lógicas, como por exemplo: a Lógica *Paraconsistente*, a Lógica *Paracompleta* e a Lógica da Diferença.

O trabalho de BARBOSA (1998) é um referencial para iniciar esse estudo.

7.5.2. Trabalhos experimentais e de manipulação experimental

Nesse caso, recomendam-se vários trabalhos que devem manter certa proximidade com o referencial teórico sobre SS já validado.

7.5.2.1. Estudo de outros métodos e técnicas de estruturação do conhecimento no domínio da CIGeo

Para iniciar a indicação de pesquisas futuras para este caso, dando continuidade à parte mais sistemática da concitação anterior (de denso conteúdo teórico), é instrutivo parafrasear MEDEIROS (1999, p. 259), quando incentivou estudos sobre a ambigüidade que afeta textos em LN armazenados em bases de dados altamente estruturadas (bases de conhecimento). Também aqui, nesta seção, *mutatis mutandis*, repete-se a concitação dessa autora, cujo escopo estava assentado na formação ou na alimentação de BCs lingüísticos e no aprimoramento de *mecanismos de recuperação* desse tipo de *informação*.

Com respeito a esses mecanismos de recuperação, a autora em tela sugeriu que a recuperação do conteúdo informativo contido em documentos[294] poderia ser resolvida pela utilização de ferramentas como *grafos conceituais*.

Apenas comentados no bojo deste trabalho, os *grafos conceituais* não foram examinados com profundidade, como o foram na tese de MEDEIROS (1999). Tais ferramentas, apesar de não serem de surgimento recente[295], têm sido exploradas de maneiras muito variadas por cientistas e pesquisadores de Lingüística Computacional e de áreas afins. Os trabalhos sobre *grafos conceituais* de SOWA (2000) elevam este tipo de ferramentas à posição de estruturas de representação do conhecimento que poderiam muito bem ser combinadas com a UML™, p.ex., para a construção de taxinomias no âmbito da CIGeo.

---

[294] Em BDs isoladas ou a ocupar nós distribuídos na rede mundial, como bibliotecas digitais (mapotecas digitais).

[295] Conceito similar às redes semânticas, que remontam a QUILLIAN (1968).

7.5..2.2. Estudo sobre a inserção de informação contextual nas ontologias, no domínio da CIGeo

Outra continuação de trabalho que se pode sugerir nessa linha é verificar a inserção de informação contextual em estruturas ontológicas que representem ou explicitem um fenômeno de natureza geográfica. Nesse caso, como já concluído, a introdução de informação contexttual só se justifica em estudos de avaliação de SS[296] entre ontologias distintas e não para uma ontologia isolada.

RODRÍGUEZ (2000) constitui um importante referencial teórico sobre o tema.

7.5..2.3. Estudos de conflito de representação (bancos de dados *vs.* ontologias)

Um outro trabalho sobre SS que poderia ser realizado, no âmbito da integração de sistemas de informação, é o de descobrir um mecanismo capaz de produzir um mapeamento consistente, que não cause conflito de representação entre as classes de entidades definidas nas ontologias, que estão num alto nível de abstração (modelo conceitual), e as mesmas classes esquematizadas em níveis mais baixos de abstração (nível de dados ou modelo operacional).

Um caso de conflito foi aqui apurado para a classe TOPONÍMIA do MC da folha Faxinal.. Na ontologia *ad-hoc*, como a carregada no PRONTO©, TOPONÍMIA não foi considerada como classe de entidades espaciais, pelos motivos apresentados no subitem 6.3.1.2.

A razão dessa proposta de trabalho funda-se na necessidade de alterar para bases mais cognitivas a ainda predominante tecnologia de bancos de dados para adquirir, processar e armazenar um incalculável e precioso acervo de dados disseminados pela *sociedade da informação*. Tais acervos estão estruturados segundo padrões caracterizados por uma semântica implícita. Contudo, apesar de ainda estar longe o dia em que se poderá obter desempenho e perenidade[297] compatíveis em níveis tão altos de abstração como nos modelos conceituais e ontologias, de semântica bem mais explícita que os esquemas de bancos de dados, como se revelou na revisão de literatura, a tendência dessa evolução é irreversível.

Esse mecanismo de mapeamento proposto aponta na direção da interoperabilidade de sistemas de informação. No caso em tela, trata-se da interoperabilidade entre os sistemas que estão em uso e já não correspondem mais às necessidades de interação homem-máquina, mas que armazenam o item mais caro no ciclo de vida dos SIs – os dados -, e a-

---

[296] RODRÍGUEZ (2000, p.101) denomina essa relação de SS cruzada.

[297] Preservação dos dados em suportes de memória estáveis (V. PREVAYLER™, subitem 4.3).

queles protótipos de sistemas que, apesar de ainda estarem nos seus primórdios[298], já esboçam altíssima interação (comunicação) com o usuário e que necessitam de matéria-prima mais sutil – informação[299] - do que dados de pouco conteúdo semântico.

7.5.2.4. Estudo de integração da similaridade espacial e da semântica

As primeiras pesquisas na área[300] de avaliação de similaridade entre entidades espaciais começaram em meados da década de 90 e contemplaram mais os aspectos geométricos. Como neste trabalho não se tratou das propriedades geométricas dessas entidades, mas das propriedades de ordem cognitiva que a avaliação da similaridade semântica suscita em relação ao domínio espacial, ou seja, dos aspectos semânticos dos termos que se referem às classes de entidades espaciais e de suas relações hierárquicas e meronímicas, pode-se também propor um trabalho cujo objetivo a se atingir seja o da integração entre esses dois ramos da avaliação de similaridade entre entidades espaciais (o geométrico e o semântico), como tem sido concitado em todas as teses emanadas dessa linha de pesquisa.

O ponto de partida para um trabalho integrador entre os enfoques espacial e o semântico da similaridade entre classes de entidades espaciais é o artigo de BRUNS (2001). Em língua portuguesa, existe a tese de doutorado de PAIVA (1998)[301].

Em geral, o foco de trabalhos como o das duas literaturas citadas *ut retro* concentra-se em bases de dados geográficos e na sua capacidade de representar as diversas visões da realidade geográfica, em vez de tratar dos elementos semânticos limitados apenas pelo desenvolvimento linear do texto (termos, toponímia, frases) numa folha de papel. Um mapa, p.ex., possui características multidimensionais [MARTINELLI (1991) e PRADO (2001)].

No domínio das entidades e relações genuinamente espaciais, entram em jogo variáveis visuais para mensurar os elementos gráficos pontuais, lineares e planares, como: posição, tamanho, valor, granulação, cor, orientação e forma (V. *Neografia* no subitem 1.5.1).

Para conhecer essas entidades formadas pelas primitivas euclidianas (ponto, reta e plano), métricas da Engenharia Cartográfica como escala de representação, precisão gráfica e relações topológicas são empregadas para simbolizar a diversidade de dados espaciais coletados de uma área, às vezes de variadas fontes, num suporte concreto (plástico, papel, etc.) ou num suporte magnético (disco, fita, etc.). O primeiro suporte (tradicional mas ainda em uso) impõe estabilidade ao seu conteúdo informativo, já que o seu domínio de difusão

---

[298] No papel ou até mesmo nos sonhos...
[299] V. visualização preliminar no subitem 3.2.2.1.
[300] Relatadas em RODRÍGUEZ (2000, p. 2).
[301] Para consultar a tese na íntegra: http://www.dpi.inpe.br/teses/miro

depende do monopólio do conhecimento das técnicas de produção do seu proprietário autoral. O segundo suporte, em franca expansão, imprime uma tal maleabilidade ao seu conteúdo informativo e desencadeia um sem-número de formas de difusão, que rompem o monopólio do conhecimento do seu produtor.

Não importando o meio em que uma mapa esteja representado, um estudo desse tipo deve ter como objetivo geral, numa primeira fase, levantar os requisitos de integração que garantam uma consistência adequada entre as diversas bases de dados geográficas perante usuários que almejem respostas de natureza espacial para as suas buscas (*queries*). Para isso, o conceito de *similaridade espacial* é essencial. Para se ter uma idéia geral desse conceito, basta imaginar três imagens, mantendo-se uma fixa (referencial). Das outras duas, a que mais transformações sofrer para se assemelhar em forma e estrutura da que foi escolhida como referencial, será a menos similar (BRUNS, 2001).

A etapa mais difícil dessa proposta seria a de sintetizar ou interpretar os resultados obtidos pelas pesquisas sobre ***similaridade espacial*** e ***similaridade semântica*** (SS). Pelo que se reviu de literatura nesta pesquisa, ainda não houve uma tentativa nesse sentido, mas esforços concentrados de trabalho numa linha ou noutra, com todas as limitações inerentes aos métodos de experimentação existentes[302], já podem constituir um material de investigação suficiente para desencadear tal estudo. Uma conclusão sobre uma possível insuficiência de resultados já seria uma contribuição, todavia o que se almeja realmente é o Santo Graal da Metodologia Científica, em que a CIGeo exerceria um papel protagonista, ao propiciar a complementação dos métodos quantitativos de avaliação de similaridade espacial com os métodos de natureza qualitativa da SS de classes de entidades espaciais, preparando o caminho para os SIGAIAs (sistemas de informações geográficas apoiados pela IA).

7.5.2.5. Estudo sobre a criação de uma linguagem de representação de ontologias

Tanto na tese de RODRÍGUEZ (2000, p. 137) como nesta tese, foram produzidos protótipos para automatizar a construção de uma ontologia e calcular a SS entre classes de entidades espaciais nelas definidas. O protótipo da primeira autora (MSS) foi implementado na LTP C++™, enquanto o PRONTO© foi implementado em *Java*™, dentro das peculiaridades metodológicas de cada estudo realizado. Tanto o MSS como o PRONTO© são protótipos de sistemas que foram implementados com LTPs de domínio público na ausência de LROs.

Um trabalho que se recomenda dentro dessa perspectiva de implementar uma estrutura de conhecimento compartilhado é produzir uma ferramenta de modelagem de ontologias.

Essas ferramentas ou LRO (linguagem de representação de ontologias) constituem sinais de que vem surgindo com a consolidação gradual de um novo campo das engenharias: a ***Engenharia de Ontologias***, ou a engenharia caracterizada por ser um conjunto de técnicas e métodos destinados a adquirir conhecimentos de domínios bem genéricos do mundo. O objetivo específico que essa nuper-surgida área parece catalisar da tendência atual verificada em ambientes acadêmicos e mesmo de mercado de *software* é produzir ferramentas de modelagem de ontologias.

De acordo com RUSSELL (1995), no âmbito das engenharias imbricadas sobre a engenharia de sistemas, o *engenheiro do conhecimento* é aquele profissional que investiga um domínio de conhecimento específico, determina[303] quais são os conceitos de interesse no domínio, depois de um extenso contacto com um grande número de casos, e estabelece qual é a LRC mais adequada para formalizar esses conceitos, não fazendo parte do perfil desse profissional a característica de ser um experto na operação de uma LRC.

Este conjunto de atribuições do engenheiro de conhecimento parece coincidir com os encargos que o seu provável colega, *engenheiro de ontologias,* irá desempenhar, quando a Engenharia de Ontologias, efetivamente, ocupar o seu espaço.

Dependendo da acomodação entre essas engenharias, poderá ocorrer uma subdivisão entre as atribuições do engenheiro de conhecimento com o de ontologias ou uma denominação de profissional poderá prevalecer e a outra se extinguir.

7.5.2.6. Estudo de comparação entre ontologias criadas por linguagens e ferramentas da OO e ontologias criadas por linguagens e aplicativos *ad-hoc*

O trabalho de desenvolver uma LRO (ou LRC, mais genericamente) pode tornar-se inviável em alcance, podendo-se sugerir como alternativa uma comparação entre as várias LROs e aplicativos de construção de ontologias existentes, visualizando-se dois objetivos imediatos, que, alcançados, muito contribuiriam para a Ciência da Informação e para a CI-Geo, quais sejam:

- Obter parâmetros de avaliação da funcionalidade e do desempenho para essas linguagens e aplicativos, no que tange à sua ***expressividade*** e à sua ***representatividade***[304];

---

[302] Por questões de coerência, nesta tese, só se pode falar de limitações dos métodos de avaliação da SS.
[303] Na prática, ele é o entrevistador dos especialistas que lhe fornecerão os conceitos.
[304] Consultar esses conceitos grifados no subitem 3.2.2.3.

- Avaliar a SS numa estrutura construída com as linguagens ou ferramentas julgadas as mais promissoras e comparar os resultados com os obtidos pelo MSS, pelo PRONTO© ou qualquer outro protótipo implementado por uma LTP (OO).

Quanto ao último objetivo, o pesquisador poderia lançar mão de um *corpus* já aferido[305] para contar com um referencial comum de comparação e para adiantar o seu cronograma.

Essas considerações suplementares se fazem necessárias, para situar genericamente o pesquisador que se propuser a trabalhar nesse campo emergente, em meio às peculiaridades que nesta revisão de literatura se apurou, servindo-lhe como subsídio para começar a traçar o seu plano de pesquisa.

7.5.2.7. Estudo sobre automação na seleção de termos para compor uma ontologia

Um trabalho que contribuiria muito para a formalização de ontologias em ambiente computacional seria o estudo de um módulo de pré-carga de termos referentes a classes de entidades espaciais.

Tal módulo economizaria tempo na etapa mais demorada da formalização de uma ontologia, justamente a seleção dos elementos portadores de significado – os termos.

Além disso, dependendo do propósito da pesquisa (automação ou semi-automação na seleção), o ganho em confiabilidade na construção da ontologia seria indubitavelmente a maior contribuição de um trabalho de pesquisa dessa ordem.

Com respeito ao propósito mencionado *ut supra*, cabe esclarecer o que se pretende dizer com uma pré-carga ou seleção de termos por meios automatizados ou semi-automatizados.

No caso da *pré-carga automática*, supõe-se a completa seleção, sem qualquer intervenção do elemento humano no processo, o que equivale a dizer que a pesquisa deverá voltar-se para mecanismos de inferência da IA que estabeleçam um padrão heurístico aceitável de busca de termos[306]. Esta proposta pode ser o berço de muitos outros problemas ligados ao PLN e o pesquisador poderá explorar pequenas porções desse repertório problemático que se poderá formar. A realização de uma pesquisa desse nível pode avançar para o território de estudos de natureza exploratória.

No caso da *pré-carga semi-automática* ou seleção com supervisão do operador, o trabalho de pesquisa pode tornar-se mais viável e adequar-se a um estudo-de-caso experimental, pela interação que se operará entre o processo de apresentação gradual dos resultados

[305] Definições e axiomas já carregados nas ontologias *ad-hoc* desta tese e de RODRÍGUEZ (2000), p.ex.
[306] Resultados de pesquisas sobre *generalização cartográfica* podem oferecer algum suporte nesse estudo proposto.

pelo protótipo e a decisão de aceitá-los ou não pelo operador. Aparentemente mais lento este processo, não há como se prever se realmente o totalmente automatizado superaria o semi-automatizado em desempenho, já que os mecanismos de inferência do primeiro poderiam pender para problemas NP, dependendo dos algoritmos de busca e da abrangência do domínio de conhecimento escolhidos.

Em ambos os casos, *técnicas de engenharia reversa* (V. glossário) aplicadas aos modelos conceituais do *corpus* poderiam ser úteis para derivar os termos distribuídos pelas estruturas hierárquicas e de agregação do modelo, o qual poderia vir incompleto em matéria de documentação, como normalmente acontece.

Antes de passar ao próximo tópico, cabe relembrar um ponto importante, tocado no subitem 1.5.3 (p. 55), que trata da melhoria de um agente inteligente ARS (agente reflexo-simples). Os sete últimos subitens vão ao encontro das metas lá esboçadas.

7.5.2.8. Estudos de extensão

Como foi observado anteriormente (subitem 6.4.2), algumas alternativas experimentais não puderam ser verificadas.

Neste subitem, sugerem-se estudos-de-caso que ampliem as conclusões tiradas na presente pesquisa, particularmente quanto ao instrumento de verificação de hipóteses representado pelo PRONTO©.

Dentro dessa mesma linha de trabalho, outras alternativas estatísticas (subitem 6.4) poderiam ser tentadas, especialmente a relacionada ao *planejamento em blocos* de MOORE (2000, p. 151), que aumentaria o controle sobre o efeito de variáveis intervenientes ou ocultas ao experimento, capazes de produzir diversos efeitos sobre outras variáveis ou até mesmo prejudicar os resultados, caso introduzam uma tendenciosidade não-detectável.

Uma outra contribuição poderia ater-se ao descobrimento das variáveis intervenientes, que não foram identificadas neste estudo. Nesse caso, estender os ensaios pretendidos no subitem 6.4.2, em razão de alguns resultados[307] apresentados no subitem 6.4.1, traria para apreciação o impacto produzido por uma variável hipotética, identificada com base racional ou até mesmo intuitiva[308] pelo pesquisador. Por exemplo, esta variável interveniente poderia ser o *balanceamento da árvore n-ária*. A metodologia de pesquisa poderia orientar o pesquisador a fundamentar a existência desta variável, pela comprovação dos seus efeitos sobre as outras variáveis de pesquisa.

---

[307] No caso de PÂNTANO e ÁREA DE LAZER.

[308] Afinal, a pesquisa é de fundo cognitivo.

Um estudo desse nível seria capaz de estabelecer três tipos de efeito dessa variável interveniente sugerida[309] sobre as variáveis de pesquisa identificadas no subitem 5.3.1:

- O *balanceamento da árvore n-ária* não exerceria qualquer efeito na relação originalmente estabelecida para a SS (dependente), internodalidade (independente) e grau de superposição das feições (independente);
- O *balanceamento da árvore n-ária* exerceria um efeito significativo sobre relação originalmente estabelecida para a SS (dependente), internodalidade (independente) e grau de superposição das feições (independente), eliminando esta relação;
- O *balanceamento da árvore n-ária* exerceria uma espécie de efeito (de que natureza?) na relação originalmente estabelecida para a SS (dependente), internodalidade (independente) e grau de superposição das feições (independente), enfraquecendo esta relação.

Como se percebe da interpretação que se sugere sobre esses efeitos, não apenas sólidos conhecimentos de Estatística serão necessários, mas também suficiente experiência do pesquisador sobre o domínio científico do problema gerado. Este último fator é que faz um estudo-de-caso tornar-se suficientemente realista e, conseqüentemente, convincente na produção de informação útil para expandir as fronteiras do conhecimento.

Para esses casos particulares de interpretação de efeitos de variáveis, SIEGEL (1988, p. 254) oferece o recurso de cálculo denominado correlação parcial de Kendall por postos (*Kendall partial rank-order correlation*), disponível no aplicativo estatístico SPSS™.

Assim, o verdadeiro alcance desse estudo-de-caso poderá ser aquilatado pelos novos resultados que pesquisas como esta, aqui sugerida, entregarem à comunidade científica, ao permitir a exploração do mesmo problema de pesquisa por vários experimentos e em diferentes contextos.

7.5.2.9. Estudo sobre a introdução das tecnologias de *geoprocessamento* no campo do Planejamento e das Aplicações Militares

Ao finalizar o trabalho, é de bom alvitre produzir alguns apontamentos sobre o desenvolvimento de uma *metanóia* (nova mentalidade) sobre SIG para as Forças Armadas brasileiras, já que para organizações civis, menos dependentes do binômio hierarquia-disciplina para se manterem, o problema de harmonizar o argumento de autoridade com o científico é mais simples de se resolver.

---

[309] O pesquisador poderá descobrir outras.

Quando se discutiu a noção de integração e recuperação de informação na área de SIG, apenas se levou em consideração a área acadêmica. Mas quando se discute SIG para uma organização[310] como uma Força Armada (FA), é preciso reorientar o esforço de pesquisa em função da expectativa da política do órgão de integração dos três comandos – agora, o Ministério da Defesa. Perguntas fundamentais devem ser formuladas e respondidas pela assessoria responsável pela Política de Ciência e Tecnologia (C&T) desse ministério. Entre essas perguntas, assinalam-se:

- O que se espera de um SIG como coadjutor na tomada de decisão, quando o ***terreno***[311] é o elemento central da missão?
- Os componentes do escalão de execução da política de C&T (órgãos de direção geral) estão ambientados com a ontologia (rede conceitual) do *Geoprocessamento*? Indo além: os integrantes desse escalão possuem conhecimentos suficientes para produzir questões nesse campo?
- Onde se situa a fronteira entre o argumento de autoridade e o científico, quando o assunto for SIG?
- O órgão do qual emana a política de C&T na área de defesa do governo está preparado para pagar o preço de esvaziar o argumento de autoridade diante do argumento cientificamente sustentado, nesse caso específico?

Antes de reorientar o esforço de pesquisa no *Geoprocessamento*, as respostas a estas questões são essenciais para criar uma expectativa objetiva da FA como utente dessa TI.

Este trabalho seria de forte conotação qualitativa e, apesar de essencial para reorientar a política de C&T em bases mais objetivas e coerentes com a escassez de recursos públicos, talvez seja extemporâneo, em razão da prioridade de outros assuntos mais prementes para a atenção do poder político, da recente criação do Ministério da Defesa, ainda se acomodando às peculiaridades de cada FA e, no que tange à escolha de um assunto desse nível, da própria polêmica embutida em cada uma das perguntas formuladas. Pontos delicados, ligados a aspectos estruturais das armas e dos quadros teriam que ser tratados; p.ex: apesar do recente surgimento do campo da C&T na Doutrina Estratégica de Poder, no EB ainda não há o posto máximo na hierarquia para o Quadro de Engenheiros Militares (QEM), o que traria mais prestígio e poder de decisão aos quadros técnicos desta FA.

---

[310] Não se pretendeu referir-se a uma Força Armada como *instituição* pelos aspectos subjetivos que o termo suscita no públlico interno dessa organização, às vezes com pesada e viciada carga de subjetividade.

[311] Aqui, num sentido muito abrangente, significa: superfície terrestre, espaço aéreo e corpos d'água.

Esta tese tem compromissos tanto com a comunidade científica civil, como também com a área de C&T militar (Exército Brasileiro), em que são desejáveis, para se conquistar objjetividade e reais ganhos profissionais para o estudioso (civil ou militar) de Planejamento e Apllicações Militares do séc XXI, um genuíno espírito deontológico do escalão de decisão que abrir temas de pesquisa nesta área, assim como capacitação, sobriedade e coragem moral do pesquisador militar que optar por esta linha de trabalho.

A capacitação está ligada a um razoável preparo na área de ciências sociais aplicadas e desejável formação em Altos Estudos Militares. A coragem moral será necessária para preparar o pesquisador para o natural risco de desgastes que vivenciará, porque tocar nesses assuntos talvez seja temerário, mas não inoportuno.

Um trabalho que se proponha a analisar os aspectos comentados nos parágrafos anteriores deste subitem pairaria num nível muito complexo da realidade, visto que o mundo de estudo do pesquisador seria fortemente afetado pelos movimentos de conveniência política e das várias nuanças do relacionamento e limitações humanas. É a ***esfera política*** do desenvolvimento de sistemas de FURLAN (1998, p. 7-11) ou *camada organizacional*.

Num nível de abstração menor, na ***esfera matemática*** de FURLAN (1998, p. 7) ou *camada funcional*, cabe propor um outro trabalho, de natureza menos complexa que o anterior. Nesse nível, é possível realizar esforços de pesquisa para formar o núcleo de classes integradoras dos negócios ligados à política de defesa (preservação da integridade territorial do Brasil, p.ex.) em matéria de informação geográfica. Esse núcleo capitalizaria o que se almeja para uma FA em termos de **integração** e **recuperação precisa de informação**, pólos às vezes antitéticos na construção de um SIG de características ontológicas, como relata RODRÍGUEZ (2002): "Mais precisão geralmente implica menor poder de recuperação e vice-versa".

Na camada mais interna desse modelo organizacional encontra-se a ***esfera filosófica*** ou *camada de dados*. É dessa camada que deveria partir o esforço de desenvolvimento de um SIG ou de qualquer SI, mas normalmente ocorre o inverso, o que desvirtua a interpretação acerca do verdadeiro estado dos fatos. Seguindo-se regras racionais e já amadurecidas para trabalhar nessa dimensão, é possível obter uma visão do MR tal como ele realmente é e não pelo que sugerem as camadas mais superficiais do modelo. A boa notícia é que os quadros técnicos do Serviço Geográfico do Exército, p.ex., já identificaram os conceitos e suas relações, com o intuito de modelar todo o espaço geográfico brasileiro (EGB), em todas as dimensões fisiográficas (oito temas).

O passo seguinte seria o de identificar as operações do nível superior (esfera matemática), como já proposto. Daí, de posse de uma ontologia do EGB para uma FA como o Exército, em que as operações (processos) e os dados estariam explicitados em classes de entidades espaciais inter-relacionadas, só restaria criar um modelo organizacional que se ajustasse à realidade e não inverter todo o processo e sujeitar um sistema de informação ao enfoque superficial e subjetivo da camada mais externa: a organizacional. Eis o porquê de já se ter declarado neste subitem que um trabalho no nível organizacional seria oportuno, visto que o principal, no nível conceitual, já está pronto e o do nível seguinte (funcional) necessitaria da contribuição de um trabalho de acabamento, como já proposto.

8. REFERÊNCIAS BIBLIOGRÁFICAS

1 ABRANTES, Paulo C.C. (Org.). Introdução. In : MARTINS, D.C., BERNARDES, D., DEL NERO, H.S., TEIXEIRA, J.F., OLIVEIRA, M.B., GONZALEZ, M. E .Q., CRUZ, W.B., Paulo César Coelho Abrantes (Org.). **Epistemologia e cognição**. Brasília : Ed. Universidade de Brasília, 1994. p. 9-23.

2 ACCIOLY, R. L. *et al.* **Perspectiva histórica da inteligência artificial**. Brasília : Faculdade Alvorada de Sistemas de Informação. 1999. 58p.

3 ALCOFORADO, P. **Gottlob Frege: Lógica e Filosofia da Linguagem**. São Paulo : Ed. da USP. 1978. 158p.

4 ALMEIDA, Luís Fernando Barbosa. **A Metodologia de disseminação da informação geográfica e os metadados**. Tese de Doutorado. Centro de Ciências Matemáticas e da Natureza – UFRJ. Rio de Janeiro, 1999.

5 ALMEIDA, Napoleão Mendes. **Gramática metódica da língua portuguesa**. São Paulo : Ed. Saraiva. 1994. 698p.

6 ALSTON, William P. **Filosofia da linguagem**. Rio de Janeiro : Fahar Editores. 1973. 155p.

7 ANDRÉ, Albert. ***L'expression graphique : cartes et diagrammes***. Paris : Masson. 1980. 223p.

8 BARBOSA, Marcelo Celani. **As lógicas: as lógicas ressuscitadas segundo Luiz Sérgio Coelho de Sampaio**. São Paulo : Makron Books do Brasil Ltda. 1998. 103p.

9 BÄHR, Hans-Peter, SCHWENDER, Anita. *Linguistic confusion in semantic modelling*. In : *XVIII INTERNATIONAL CONGRESS FOR PHOTOGRAMMETRY AND REMOTE SENSING*. Vol. XXXI, *Part* B6, 1996, Viena. **Anais...** Viena : Karl Kraus, Diretor do Congresso, 1996. p. 7-12.

10 BÄHR, Hans-Peter (baehr@ipf.uni-karlsruhe.de) **Remessa do *trabalho "The power of the links"***, apresentado no simpósio anual da ISPRS – Amsterdã - 2000. 20 out. 2000. Enviada às 22h. 45 min. Mensagem para : Paulo César Rodrigues Borges (pcrborges@tba.com.br)

11 BARRETO, A. S. **Variação craniana e genética de *tursiops truncatus* (*delphinidæ, cetacea*) na costa atlântica da América do Sul**. Rio Grande, 2000. 122f. Tese (Doutorado em Oceanografia Biológica) - Universidade Federal do Rio Grande, Departamento de Oceanografia, FURG.

12 BEHRENS, Clifford *et al.* ***The geospatial interoperability problem : lessons learned from building the geolens prototype.*** Capturado em 1º jul. 2000. *Online*. Disponível na Internet : http:// www. ncgia. ucsb. edu/ cont/ interop9// program/ papers/ behrens/ behrens.html

13 BERNHARDSEN, Tor. ***Geographic information systems***. Arendal (Noruega) : Viak IT. 1982. 318p.

14 BERTIN, Jacques. ***Semiologie graphique : les diagrammes, les réseaux, les cartes***. Paris : Guathier-Villars. 1967. 431p.

15 BESSE, J-M., BOISSIÈRE, A. ***Précis de philosophie***. Paris : Nathan. 1998, p. 52-53.

16 BLASER, A.D., SESTER, M., EGENHOFER, M.J. ***Visualization in na early stage of the problem solving process in GIS***. Capturado em 28 fev. 2002. *Online*. Disponível na Internet : http://www.spatial.maine.edu/~max/RJ34.html

17 BÖHM, Corrado, JACOPINI, Guiseppe. *Flow diagrams: Turing machines and languages with only two formation rules*. ***Comm. of the ACM 9***, Nova Iorque, p. 366-371, maio 1966.

18 BRANDÃO, A. A. F. **Uma introdução à engenharia de ontologias no contexto da *web* semântica**. Capturado em 13 maio. 2003. *Online*. Disponível na Internet : http: //www. teccomm. les. Inf. puc-rio.br

19 BORGES, P. C. R. **Estruturas de dados para armazenamento de modelos digitais do terreno**. Rio de Janeiro, 1993. 434f. Dissertação (Mestrado em Sistemas e Computação) - Instituto Militar de Engenharia.

20 BORGES, P. C. R. ***A recuperação da informação geográfica e os metadados***. Capturado em 10 ago. 2002. *Online*. Disponível na Internet : http: //pcrb. tripod. com. br/pcrbking/metadados.pdf

21 BRASIL. Ministério da Defesa. Exército Brasileiro. Escola de Instrução Especializada. Curso de Topografia. **Apostila de Fotogrametria** : Topologia. Rio de Janeiro, 1985. 185p.

22 BRASIL. Ministério da Defesa. Exército Brasileiro. Estado-Maior do Exército. EGGCF. **Manual Técnico T 34-700** : Convenções Cartográficas (Primeira Parte : Norma para Emprego dos Símbolos e Segunda Parte : Catálogo de Símbolos e Convenções Cartográficas). Brasília, 1998a. 127p.

23 BRASIL. Ministério do Planejamento e Orçamento. Comissão Nacional de Cartografia. Comitê Especializado para Estudo do Padrão de Intercâmbio de Dados cartográficos digitais (CEPAD). **Atas de reuniões do CEPAD** de 19 jun. 1997 a 11 nov. 1998b.

Estabelecimento de um padrão que oriente o intercâmbio de dados cartográficos digitais no âmbito das organizações governamentais produtoras desses dados. 110 f. impressas.

24 BRASIL-NETO, J. P. ***Neurofisiologia básica***. Capturado em 10 mar. 2002. *Online*. Disponível na Internet : http: //www. geocities. com/CollegePark/Classroom/1182/

25 BRITO, Nacho. ***Prevalencia : otro enfoque a la persistencia***. Capturado em 04 abril. 2003. *Online*. Disponível na Internet : http: //www. tres-software. com/ articulos/ prevayler.htm

26 BRUNS, H. T., EGENHOFER, M. ***Similarity of spatial scenes***. Capturado em 30 março. 2001. *Online*. Disponível na Internet : http://www.spatial.maine.edu/~max/RC23.html

27 BURROUGH, P.A. ***Principles os geographical information systems for land resources assessment***. Nova Iorque : Clarendon Press – Oxford. 1987. 193p.

28 CÂMARA, G. Modelagem semântica : compreendendo as diferenças entre sistemas de *geoprocessamento*. **Revista InfoGEO**, Curitiba, p. 41-42, nov./dez. 1998.

29 CÂMARA, G. Mapas são dados, não desenhos! **Revista InfoGEO**, Curitiba, p. 33-34, jan./fev. 1999.

30 CÂMARA, G., DAVIS, C., MONTEIRO, A. M. V. (Org.). **Introdução à ciência da geoinformação**. Capturado em 13 jan. 2002. *Online*. Disponível na Internet : http: //www. dpi. inpe. br/gilbrto/livro/introd/

31 CARDOSO, Sílvia Helena (sh@nib.unicamp.br) **Interdisciplinaridade, multidisciplinaridade e transdisciplinaridade**. 26 jun. 2001. Enviada às 19h. 56min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

32 CHAUÍ, Marilena. **Convite à Filosofia**. São Paulo : Ed. Ática. 1998. 440p.

33 CHOMSKY, Noam. Linguagem e mente. Brasília : Editora Universidade de Brasília. 1998. 83p.

34 CILIATO, M. Z. H. **Escolha da aptidão profissional utilizando ferramenta computacional**. Frederico Westphalen : Universidade Regional Integrada do Alto Uruguai e das Missões - Departamento de Engenharia e Ciências da Computação. 2000. 105p.

35 CLUL - Centro de Lingüística da Universidade de Lisboa. ***Computação do conhecimento léxico-gramatical***. Capturado em 18 jun. 2002. *Online*. Disponível na Internet http://www.clul.ul.pt/sectores/projecto_wordnet.html

36 COSTA, C. F. **Filosofia da linguagem**. Rio de Janeiro : Jorge Zahar Ed. 2002, 60p.

37 COUTO, Hildo Honório. **Uma introdução à semiótica**. Rio de Janeiro: Presença Edições. 1983. 162p.

38 CRANEFIELD, S., PURVIS, M. ***UML as na ontology modelling language***. Capturado em 23 jul. 2002. *Online*. Disponível na Internet :http: //citeseer. nj. nec. com/cranefield99uml. html

39 CRUZ, W. B., BERNARDES, D., MARTINS, D. A. C. A inteligência dos computadores: sugestão de metodologia de abordagem do problema. In : MARTINS, D.C., BERNARDES, D., DEL NERO, H.S., TEIXEIRA, J.F., OLIVEIRA, M.B., GONZALEZ, M. E .Q., CRUZ, W.B., Paulo César Coelho Abrantes (Org.). **Epistemologia e cognição**. Brasília : Ed. Universidade de Brasília, 1994. p. 219-226.

40 CURRÁS, Emília. ***Tesauros*, linguagens terminológicas**. Brasília : Instituto Brasileiro de Informação em Ciência e Tecnologia. 1995. 286p.

41 DAGHLIAN, Jacob. **Lógica e álgebra de Boole**. São Paulo : Editora Atlas S.A. 1995. p.18.

42 DAHLBERG, Ingetraut. Teoria do conceito. **Revista Ciência da Informação**, Rio de Janeiro, 7(2), p. 101-107. 1978.

43 D'ALGE, J. C. L., GOODCHILD, M. F. **Generalização Cartográfica, Representação do Conhecimento e SIG.** Capturado em 29 jul. 2002. *Online*. Disponível na Internet : http: //www. dpi. inpe. br/~julio/sbsr96. Pdf

44 DANIEL. W. W. ***Applied nonparametric statistics***. Boston : Houghton Mifflin Company. 1978. 503p.

45 DAVIS, Clodoveu A. J. Generalização em GIS. **Revista InfoGEO**, Curitiba, p. 40-41, jan./fev. 1999.

46 DAWN, J. Wright, GOODCHILD, Michael F., PROCTOR, James D. ***Demystifying the persistent ambiguity of GIS as "tool" versus "science"***. Capturado em 12 jun. 2000. *Online*. Disponível na Internet : http: //dusk. geo. orst. edu/annals. Html

47 DCCUT – DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO DA UNIVERSIDADE DO TEXAS. ***Boolean and vector space retrieval models***. Capturado em 24 jun. 2002. *Online*. Disponível na Internet http: //www. cs. utexas. edu/users/mooney/ir-course/slides/IRmodels.ppt

48 DEITEL, H. M., DEITEL, P. J. **Java : como programar**. Porto Alegre : Bookman Cia. Ed. 2001. 1.201p.

49 DFJUG – JAVA USER'S GROUP do DF. ***A prevalência de objetos – PREVAYLER***. Boletins eletrônicos DFJUG nº 226, 227 e 228, Brasília, 2002, Brasília. *Online*. Disponível na Internet http: //www. dfjug. org/ portugues/ boletins/ Ano02.htm

50 DIAS, M . G . B. **Organização do conhecimento utilizado na manutenção de *software***. Brasília, 2001. Apresentação visual. Dissertação (Mestrado em GC e TI). Universidade Católica de Brasília.

51 DIJKSTRA, Edsger W. *Go to: statement considered harmful.* ***Comm. of the ACM 11***, Nova Iorque, p. 147-148, março 1968.

52 DUCROT, Oswald e TODOROV, Tzvetan. **Dicionário enciclopédico das ciências da linguagem**. São Paulo : Editora Perspectiva. 1972. 339p.

53 EGENHOFER, M., MARK, D. *Naive Geography*. In : A. Frank and W. Kuhn (eds.). ***A theoritical basis for geographic information systems***. International Conference COSIT'95. Semmering (Austria) : Spring-Verlag, 1995. p. 1-14.

54 EGENHOFER, M. J., KUHN, W. (a) ***Interacting with geographic information systems***. Capturado em 10 jul. 2001. *Online*. Disponível na Internet : http: //www. spatial. maine. edu/~max/max. html

55 EGENHOFER, M. J. (b) ***Interoperability theory***. Capturado em 15 jul. 2001. *Online*. Disponível na Internet : http://www.spatial.maine.edu/~max/max.html

56 EYSENCK, M. W., KEANE, M. T. **Psicologia cognitiva : um manual introdutório.** Porto alegre : Ed. Artes Médicas Ltda. 1994. 490p.

57 FELBER, Helmut, PICHT, H. ***Metodos de terminografia y principios de investigacion terminologica***. Madri : Instituto "Miguel de Cervantes" - CSIC, 1984. 254 p.

58 FETI - FACULDADE DE ENGENHARIA E DE TECNOLOGIA DA INFORMAÇÃO DA UNIVERSIDADE DA CAROLINA DO SUL. ***Vector space systems***. Capturado em 24 jun. 2002. *Online*. Disponível na Internet http: //www. cse. sc. edu/ ~eastman/ CSCE725/ Vector%20Space%20Systems.ppt

59 FITZGERALD, Arthur Eugene, HIGGINBOTHAM, David E., GRABEL, Arvin. Engenharia elétrica. São Paulo: McGraw-Hill do Brasil. 1981. 744p.

60 FOLEY, James D., VAN DAM, Andries. ***Fundamentals os interactive computer graphics***. Califórnia : Addison-Wesley Publishing Company. 1984. 664p.

61 FONSECA, Frederico. Onde estou, para onde vou?. **Revista InfoGEO**, Curitiba, p. 18-19, nov./dez. 2000a.

62 FONSECA, Frederico, EGENHOFER, Max, BORGES, Karla A. V. Ontologias e Interoperabilidade Semântica entre SIGs. In : **II Workshop Brasileiro em Geoinformática - GeoInfo2000**. *Proceedings*. São Paulo. 2000b.

63 FPCEUC - FACULDADE DE PSICOLOGIA E CIÊNCIAS DA EDUCAÇÃO DA UNIVERSIDADE DE COIMBRA. ***Psicologia funcionalista***. Capturado em 12 jul. 2002. *Online*. Disponível na Internet : http://www.fpce.uc.pt/nucleos/niips

64 FREEMAN, James A., SKAPURA, David M. ***Neural networks : algorithms applications, and programming techniques***. Califórnia : Addison-Wesley Publishing Company Inc. 1991. 400p.

65 FURLAN, José Davi. **Modelagem de objetos através da UML : análise e desenho orientados a objeto**. São Paulo : Makron Books do Brasil Ed. Ltda. 1998. 329p.

66 GAHEGAN, Mark N. *Characterizing the semantic content of geographic data, models, and systems*. In : EGENHOFER, Max J., FEGEAS, Robin, KOTTMAN, Cliff, Michael Goodchild (Org.). ***Interoperanting geographic information systems***. Norwell : Kluwer Academic Publishers, 1999. p. 71-84.

67 GANESAN, P., GARCIA-MOLINA, H., WIDOM, J. Exploiting hierarchical domain structure to compute similarity. Capturado em 15 mar. 2002. *Online*. Disponível na Internet http: //dbpubs. stanford. edu/pub/2001-26

68 GARCIA, O. M. **Comunicação em prosa moderna**. Rio de Janeiro : FGV. 1976. 508p.

69 GOMES, Hagar Espanha. **Elaboração de *tesauro* documentário : aspectos teóricos e práticos**. Rio de Janeiro : edição da autora. 1998. 33p.

70 GONDIM, P. R. L. Redes neurais: computação convencional x neurocomputação, ferramentas e tecnologias para implementação. **Revista Militar de Ciência e Tecnologia**, Rio de Janeiro, p. 8-15, abr../jun. 1991.

71 GONÇALVES, N. V. **Modelo de recuperação de informações temáticas interrelacionadas, contidas em imagens de satélites, baseado em descritores contextuais**. Brasília, 2001. 225f. Tese (Doutorado em Ciência da Informação) - Faculdade de Estudos Sociais Aplicados, UnB.

72 GONZALEZ, M. E. Q. Um estudo cognitivo-informacional das representações mentais. In : MARTINS, D.C., BERNARDES, D., DEL NERO, H.S., TEIXEIRA, J.F., OLIVEIRA, M.B., GONZALEZ, M. E .Q., CRUZ, W.B., Paulo César Coelho Abrantes (Org.). **Epistemologia e cognição**. Brasília : Ed. Universidade de Brasília, 1994. p. 127-146.

73 GORSKI, D. P., TAVANTS, P. V. ***Logica***. México (DF) : Ed. Grijalbo. 1968. p. 38-83.

74 GUIMARÃES, A. M., CASTILHO, N. A. Algoritmos e estruturas de dados. Rio de Janeiro: LTC Ed. 1985. 216p.

75 GUINCHAT, Claire, MENOU, Michel. **Introdução geral às ciências e técnicas da informação e documentação**. Brasília : IBICT, 1994. 540p.

76 GUIZZARDI, G. **Análise de domínio e ontologias**. Capturado em 13 maio. 2003. *Online*. Disponível na Internet : http://wwwhome.cs.utwente.nl/~guizzard/MSc/cap3.pdf

77 HARRINGTON, Steven. ***Computer graphics : a programming approach***. Nova Iorque : McGraw-Hill Book Company. 1983. 448p.

78 HECHT-NIELSEN, Robert. ***Neurocomputing***. Califórnia : Addison-Wesley Publishing Company Inc. 1989. 433p.

79 HENRIQUES, A. **A categorização do conhecimento**. Capturado em 12 out. 2001. *Online*. Disponível na Internet http: //pwp. netcabo. pt/0165008002/ cogni% 202% 20acet% categoriz .htm#top.

80 HOROWITZ, E., SAHNI, S. ***Fundamentals of data structures in Turbo Pascal***. Maryland : Computer Science Press. 1989. 478p.

81 HUISMAN, Denis, VERGEZ, André. **Curso moderno de filosofia : introdução à filosofia das ciências**. Rio de Janeiro : Biblioteca Universitária Freitas Bastos, 1976. 339p.

82 HWANG, Kai, BRIGGS, Fayé A. ***Computer architecture and parallel processing***. Cingapura : McGraw-Hill International Editions. 1985. 846p.

83 HYVARINEN, Aapo. ***Principal component analysis***. Capturado em 24 jul. 2002. *Online*. Disponível na Internet http: //www. cis. hut. fi/~aapo/papers/NCS99web/node5. html 24 jul. 2002.

84 HYVARINEN, Aapo. ***Factor analysis***. . Capturado em 24 jul. 2002. *Online*. Disponível na Internet: http: //www. cis. hut. fi/~aapo/papers/NCS99web/node6. html

85 INGWERSEN, Peter. Cognitive perspectives of information retrieval interaction elements of a cognitive IR theory. ***Journal of Documentation***, Copenhague, p. 3-50, v. 52, n.1, mar. 1996.

86 IPUSP - INSTITUTO DE PSICOLOGIA DA USP. **Psicologia sensorial**. Capturado em 10 mar. 2002. *Online*. Disponível na Internet : http: //www. usp. br/ ip/ posg/ nec/ nec_disciplinas. htm

87 JANNUZZI, C. A. S. C. ***Estoque, oferta e uso da informação : reflexões sobre um recurso estratégico para o desenvolvimento do setor produtivo***. Capturado em 15 ago. 2002. *Online*. Disponível na Internet : http: //www. mdic. gov. br/ tecnologia/ revistas/artigos/ SPcamp/art02CelesteAidaJannuzzi. PDF

88 JIANG, J. J., CONRATH, D. W. *). Semantic similarity based on corpus statistics and lexical taxonomy*. In : International Conference on Computational Linguistics (ROCLING X), Taiwan, **Anais ...** p. 19-35. 1997.

89 JOHN, G. H., KOHAVI, R., PFLEGER, K. *Irrelevant features and the subset selection problem.* In : Eleventh International Conference of Machine Learning, São Francisco, **Anais ...** p. 121-129. 1994.

90 KHALFA, J. (Org.). Introdução. In : GREGORY, R., MACKINTOSH, N., BUTTERWORTH G., SCHANK, R. e BIRNBAUM, L., PENROSE, R., AROM, S., DENNET, D., SPERBER, D, KHALFA, Jean Khalfa (Org.). **A natureza da inteligência.** São Paulo : Fundação Ed. da UNESP, 1995. p. 7-17.

91 KELLER, F. S. Psicologia : problemas históricos. **Boletim de Psicologia da Sociedade de Psicologia de São Paulo**, São Paulo, p. 1-33, n. 43, 1964.

92 KIRYAKOV, A. K. *et al.* ***Ontologically supported semantic matching***. Sofia : Bulgarian Academy of Sciences. 1998. 13p.

93 KONDRATOV, A. **Sons e sinais : semiótica, cibernética, lingüística, lógica**. Brasília: Coordenada – Ed. de Brasília. 1972. 189p.

94 KREIDER, D., KULLER, R. G., OSTBERG, R. R., PERKINS, F. W. **Introdução à análise linear : equações diferenciais lineares**. Rio de Janeiro: LTC S.A. 1972. 314p.

95 KRUMHANSL, C. L. *Concerning the applicability of geometric models to similarity data: the interrelationship between similarity and spatial density.* ***Psychological Review***. Califórnia, v. 85, n. 5, p. 445-463. 1978.

96 KUHN, Werner. *Metaphors create theories for users.* In : Spatial Information Theory, European Conference COSIT '93. Lecture Notes in Computer Science, 1993, Itália. **Anais...** Nova Iorque : A. Frank e I. Campari, Editores, Springer-Verlag, 1993. p. 366-376.

97 LASER-SCAN, Inc. ***Mapping and charting***. Capturado em 28 jul. 2002. *Online*. Disponível na Internet http: //www. laser-scan. com/company/index. Htm

98 LEÃO, P. R. C. **Gestão de competências profissionais em TI** (título provisório). Brasília, 2003. Apresentação visual. Projeto de dissertação (Mestrado em GC e TI) – Universidade Católica de Brasília.

99 LEINZ, V., DO AMARAL, S.E. **Geologia geral**. São Paulo : Ed. Nacional. 1980. 397p.

100 LELLO, J., LELLO, E. **Lello universal : dicionário enciclopédico luso-brasileiro em quatro volumes**. Porto : Lello & Irmãos Editores. 1984. 1224p.

101 LEVIN, J. **Estatística aplicada a ciências humanas**. São Paulo : Ed. Harbra ltda. 1987. p. 193-265.

102 LÉVY, Pierre. ***Les technologies de l'intelligence***. Paris : Éditions La Découverte. 1990. 203p.

103 LIPSCHUTZ, S. Álgebra linear. São Paulo: Ed. McGraw-Hill do Brasil. 1972. 413p.

104 LOCKE, L. F., SPIRDUSO, W. W., SILVERMAN, S. J. ***Proposals that work : a guide for planning dissertations and grant proposals.*** Londres : SAGE Publications. 1993. 323p.

105 LOPES, Edward. **Metáfora: da retórica à semiótica**. São Paulo : Atual Editora Ltda. 1987. 112p.

106 LUGNANI, João Bosco. **Introdução à fototriangulação**. Curitiba : Universidade Federal do Paraná. 1987. 134p.

107 LUNARDI, O. A. **Modelagem do espaço geográfico brasileiro**. Brasília, CCAuEx, 17 de maio de 2001. Palestra ministrada para oficiais engenheiros militares do Exército da guarnição de Brasília, para grupos de trabalho da área de inteligência do Exército e para representantes das empresas *SulSoft* (revendedora do programa ENV ®) e da Laser-Scan (revendedora do sistema *Gothic*®).

108 McADAMS, H. T., CRAWFORD, R. W., HADDER, G. R. ***A vector approach to regression analysis and its application to heavy-duty diesel emissions***. Capturado em 11 jul. 2002. *Online*. Disponível na Internet : http: //www. ncgia. ucsb. edu/ cont/ interop9// program/papers/behrens/behrens. html

109 MALMBERG, Bertil. **A língua e o homem : introdução aos problemas gerais da lingüística**. Rio de Janeiro : Livraria Duas Cidades. 1976. 182p.

110 MARK, D. (dmarh@geog.buffalo.edu) ***Maps versus text***. 04 ago. 2002. Enviada às 13h. 03 min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

111 MARTINELLI, Marcelo. **Curso de cartografia temática**. São Paulo : Editora Contexto. 1991. 180p.

112 MASUDA, Yoneji. **A sociedade da informação como sociedade pós-industrial**. Rio de janeiro : Editora Rio. 1982. 212p.

113 MEDEIROS, M. B. B. **Tratamento automático de ambigüidades na recuperação da informação**. Brasília, 1999. 290f. Tese (Doutorado em Ciência da Informação) - Faculdade de Estudos Sociais Aplicados, UnB.

114 MILLER, George, CHARLES, W. G. *Contextual correlates of semantic similarity*. Nova Jérsei: ***Language and Cognitive Processes***, v. 6, n. 1, p. 1-28. 1991.

115 MILLER, R. L., ACTON, C., FULLERTON, D. A., MALTBY, J. ***SPSS for social scientists***. Houndmills : Palgrave Macmillan. 2002. 334p.

116 MIRANDA, M. N. **Algoritmos genéticos : fundamentos e aplicações**. Capturado em 16 jul. 2002. *Online*. Disponível na Internet http: //www. gta. ufrj. br/~marcio/genetic. html#Principais

117 MODELL, Martin (a) ***Classification overview***. Capturado em 12 ago. 2001. *Online*. Disponível na Internet : http: //www. dai-sho. com/dadmc/dadmc09. html

118 MODELL, Martin (b) (marty.modellh@dai-sho.com). ***What's the difference between characteristics and properties?*** 12 ago. 2001. Enviada às 17h. 44min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

119 MOLENAAR, Martien. *Status and problems of geographical information systems: the necessity of a geoinformation theory*. ***Journal of Photogrammetry and Remote Sensing***, Amsterdã, 46, 1991. p. 85-103.

120 MOORE, D. S. **A estatística básica e a sua prática**. Rio de Janeiro : LTC Ed. 2000. p. 142-155.

121 MORA, José Ferrater. **Dicionário de Filosofia (tomo IV)**. São Paulo : Edições Loyola. 1994. p. 892-895, p. 1.380-1.383, p.2.340, p. 2.630-2.635.

122 MOURA, Ana Clara Mourão. Cartografia temática como meio de comunicação. **Revista FatorGIS**, Curitiba, n.6, p. 25-27, jul./ago./set. 1994.

123 NEWMAN, William M., SPROULL, Robert F. ***Principles of interactive computer graphics***. Nova Iorque : McGraw-Hill Book Company. 1979. 541p.

124 NORTON, Peter. **Desvendando o PC**. Rio de Janeiro: Ed. Campus, 1989. 239p.

125 NORVIG. Peter (peter@norvig.com) ***KRLs and oriented object programming languages***. 20 fev. 2002. Enviada às 13h. 15 min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

126 ORTONY, Andrew. ***Metaphor and thought***. Nova Iorque : Editora da Universidade da Cambrígia. 1988. 501p.

127 PAIVA, J. ***Topological equivalence and similarity in multiple representation geographic database***. Maine, 1998. ?f. Tese (Doutorado em Ciência da Informação Espacial e Engenharia) - Dep. de Ciência da Informação Espacial e Engenharia, Universidade do Maine.

128 PASCUAL, Antonio F.R. ***Sistemas de información geográfica en España : un campo sin cartografiar***. Madri : Fundación Conde del Valle Salazar – IGNE, 1993.

129 PASSOS, E. L., PEREIRA, P. C. A. Redes neurais artificiais: uma introdução. **Revista Militar de Ciência e Tecnologia**, Rio de Janeiro, p. 7-11, jan./mar. 1990.

130 PENROSE, Roger. Inteligência matemática. In : GREGORY, R., MACKINTOSH, N., BUTTERWORTH G., SCHANK, R. e BIRNBAUM, L., PENROSE, R., AROM, S., DENNET, D., SPERBER, D, KHALFA, Jean Khalfa (Org.). **A natureza da inteligência.** São Paulo : Fundação Ed. da UNESP, 1995. p. 139-162.

131 PINTO, Mílton José. **Análise semântica de línguas naturais : caminhos e obstáculos**. Rio de Janeiro : Editora Forense Universitária. 1977. 125p.

132 POPPER. K. R. **Conhecimento objetivo: uma abordagem evolucionária**. Delo Horizonte : Ed. Itatiaia Ltda. 1975. 394p.

133 PRADO, A. B., BARANAUSKAS, M. C. C., MEDEIROS, C. M. B. **Cartografia e sistemas de informação geográfica como sistemas semióticos : uma análise comparativa**. Capturado em 06 nov. 2001. *Online*. Disponível na Internet http: //www. ecgraf. puc-rio. br/geoinfo2000/anais/002. pdf

134 PRADO, Hélio Gouvêa. **Restituição digital fotogramétrica**. Rio de Janeiro, 1992. 117f. Dissertação (Mestrado em Sistemas e Computação) - Instituto Militar de Engenharia, RJ.

135 PRATT, Ian, LEMON, Oliver. ***A formal semantics of cartographic representation***. Capturado em 1º nov. 2000. *Online*. Disponível na Internet : http: //www. cs. man. ac. uk/ai/oliver/mapsem. html

136 PRESSMAN, R. S. **Engenharia de *software***. São Paulo : Makron Books do Brasil Ed. Ltda. 1995. 1055p.

137 PUCSP - PONTIFÍCIE UNIVERSIDADE CATOLÓLICA DE SP. **Classificação**. Capturado em 13 jan. 2002. *Online*. Disponível na Internet : http: //www. pucsp. br/~logica/Classificacao. htm

138 QUILLIAN, M. Ross. *Semantic memory*. In : RAPHAEL, Bertram, BOBROW, Daniel G., QUILLIAN, M. Ross, EVAN, Thomas G., BLACK, Fischer, McCARTHY, John, Marvin L. Minsky (Org.). ***Semantic information processing***. Cambridge : MIT Press, 1968. p. 227-270.

139 RADA, R., MILI, H., BICHNELL, E., BLETTNER, M. *Development and application of a metric on semantic nets*. (?) : **IEEE**, v. 19, n. 1, p. 18-30. 1989.

140 RAISZ, Erwin. **Cartografia geral**. Rio de Janeiro : Editora Científica. 1969. 414p.

141 RAYCHAUDHURI, S., STUART, J . M., ALTMAN, R. B (altman@smi.stanford.edu) Remessa do trabalho ***"Principal components analysis to summarize microarray experiments : application to sporulation time series"***, publicado pela Stanford Medical

Informatics – 1999. 30 mai. 2002. Enviada às 20h. 45 min. Mensagem para : Paulo César Rodrigues Borges (pcrb@terra.com.br)

142 RAPHAEL, Bertram. *SIR : sematinc information retrieval.* In : RAPHAEL, Bertram, BOBROW, Daniel G., QUILLIAN, M. Ross, EVAN, Thomas G., BLACK, Fischer, McCARTHY, John, Marvin L. Minsky (Org.). ***Semantic information processing***. Cambridge: MIT Press, 1968. p. 33-145.

143 RESNICK, P. *Semantic similarity in a taxonomy : an information-based measure and its application to problems of ambiguity in natural language.* ***Journal of Artificial Intelligence Ressearch***, Maryland, v. 11, p. 95-130. 1999.

144 RICHARDSON, Roberto Jarry *et al.* **Pesquisa social : métodos e técnicas**. São Paulo: Editora Atlas S.A. 1999. 334p.

145 RICH, Elaine, KNIGHT, Kevin. **Inteligência artificial**. São Paulo : Makron Books do Brasil Ed. Ltda. 1993. 722p.

146 RIPS, Lance, SHOBEN, E. J. *Semantic distance and the verification of semantic relations.* ***Journal of verbal learning and verbal behavior***. Califórnia, v. 12, p. 1-20. 1973.

147 ROBERTSON, S. E. ***Computer retrieval as seen through the pages of journal of documents***. Londres : B. C. Vickery Ed., 1994. p. 119-146.

148 RODRÍGUEZ, M. Andrea T., EGENHOFER, Max, RUGG, Robert D. ***Assessing semantic similarities among geoespatial feature class definitions***. Maine : Universidade do Maine. 2000. 10p.

149 RODRÍGUEZ, M. Andrea T. ***Assessing semantic similarities among spatial entity classes***. Maine, 2000. 168f. Tese (Doutorado em Ciência da Informação Espacial e Engenharia) - Dep. de Ciência da Informação Espacial e Engenharia, Universidade do Maine.

150 RODRÍGUEZ, M. Andrea T (andrea@udec.cl) ***Doubts about conclusions and future research directions***. 15 mai. 2001a. Enviada às 16h. 15 min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

151 RODRÍGUEZ, M. Andrea T. (andrea@udec.cl) ***Sending MD model code***. 25 jul. 2001b. Enviada às 12h. 35 min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

152 RODRÍGUEZ, M. Andrea T. (andrea@udec.cl) ***Maps versus text***. 27 jul. 2002. Enviada às 18h. 30 min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

153 ROGERS, David F. ***Procedural elements for computer graphics***. Nova Iorque : McGraw-Hill International Editions. 1985. 433p.

154 RUDIO, Franz Victor. **Introdução ao projeto de pesquisa científica**. Petrópolis : Editora Vozes. 1978. 121p.

155 RUELLE, David. **Acaso e caos**. São Paulo : Ed. da UNESP. 1993. 224p.

156 RUMBAUGH, James *et al.* **Modelagem e projetos baseados em objetos**. Rio de Janeiro : Editora Campus. 1994. 652p.

157 RUMMEL R. J. ***Understanding factor analysis***. . Capturado em 24 jul. 2002. *Online*. Disponível na Internet http: //www. hawaii. edu/powerkills/UFA. HTM 30 mai. 2002

158 RUSSELL, S. J., NORVIG, P. ***Artificial intelligence : a modern approach***. Nova Jérsei : Prentice-Hall, Inc. 1995. 932p.

159 SABBATINI, Renato M.E., CARDOSO, Sílvia H. ***Interdisciplinarity and the study of mind.*** Capturado em 08 set. 2000. *Online*. Disponível na Internet : http: //www. epub. org. br/cm/n06/opiniao/interdisc_i. htm

160 SALMON, W. C. **Lógica**. Rio de Janeiro : Zahar Ed. 1973. 142p.

161 SANTAELLA, Lúcia. **O que é semiótica**. Brasília : Ed. Brasiliense S.A. 1983. 84p.

162 SANTOS, T. L . V. L.. **Algoritmos genéticos para ordenamento em sistemas de busca na *web***. Capturado em 08 mar. 2002. *Online*. Disponível na Internet : http: //www. cin. ufpe. br/~tlvls/tg/

163 SAUSSURE, F. **Curso de lingüística geral**. São Paulo : Editora Cultrix. 1975. 279p.

164 SCHANK, Roger, BIRNBAUM, Lawrence. Aumentando a inteligência. In : GREGORY, R., MACKINTOSH, N., BUTTERWORTH G., SCHANK, R. e BIRNBAUM, L., PENROSE, R., AROM, S., DENNET, D., SPERBER, D, KHALFA, Jean Khalfa (Org.). **A natureza da inteligência.** São Paulo : Fundação Ed. da UNESP, 1995. p. 77-110.

165 SEKARAN, Uma. ***Research methods for business : a skill-building approach***. Nova Iorque: John Wiley & Sons, Inc. 2000. 463p.

166 SETZER, V. W. **Bancos de dados : conceitos, modelos, gerenciadores, projeto lógico, projeto físico**. São Paulo : Editora Edgar Blücher Ltda. 1989. 289p.

167 SETZER, V. W. **Dado, informação, conhecimento e competência**. Capturado em 30 dez. 2001. *Online*. Disponível na Internet : http: //www. ime. usp. br/~vwsetzer/dado-info. Html

168 SHETH, Amit P. *Changing focus on interoperability in information systems : from system, syntax, structure to semantics*. In : EGENHOFER, Max J., FEGEAS, Robin, KOTTMAN, Cliff, Michael Goodchild (Org.). ***Interoperanting geographic information systems***. Norwell : Kluwer Academic Publishers, 1999. p. 5-30.

169 SIEGEL, S, CASTELLAN, N. J. ***Nonparametric statistics for the behavioral sciences***. New York : McGraw-Hill, Inc. 1988. 399p.

170 SILVA, Marcos Antônio da. **Análise comparativa entre os procedimentos de construção de linguagem documentária aplicada à linguagem de modelagem unificada (UML) e ao tesauro**. Brasília, 2001. 264f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Estudos Sociais Aplicados, unB, Brasília.

171 SILVA, Wagner Teixeira da. **Algoritmos para raciocínio evidencial usando funções de crença**. São Paulo, 1991. 214f. Tese (Doutorado em Informática) – Departamento de Informática, PUC, RJ.

172 SMEATON, A. F., QUIGLEY, I. ***Experiments on using semantic distances between words in image caption retrieval***. Ireland: School of Computer Applications of Dublin City University (working paper). 1997.

173 SÖRGEL, Dagobert. *The rise of ontologies or the reinventation of classification*. **JASIS**, Maryland, v. 50, n. 12, p. 1.119-1.120. 1999.

174 SAGAN, Carl. **Os dragões do éden**. São Paulo : Círculo do Livro S.A. 1977. 220p.

175 SBF - SOCIEDADE BRASILEIRA DE FISIOLOGIA. **Boletim 23(1), 1998**. Capturado em 10 mar. 2002. *Online*. Disponível na Internet : http: //www. geocities. com/sbfis2000/bol23198. htm

176 SEP - STANFORD ENCYCLOPEDIA OF PHILOSOPHY. ***Cognitive Science***. Capturado em 12 ago. 2001. *Online*. Disponível na Internet : http: //plato. stanford. edu/rentries/cognitive-science/

177 SONESSON, G. ***Iconicity in the ecology of semiosis***. Capturado em 08 ago. 2002. *Online*. Disponível na Internet : http: //www. arthist. lu. se/ kultsem/ sonesson/ LifeworldIconicity2. html

178 SOWA, J. F. ***Knowledge representation : logical, philosophical, and computational foundations***. California : Pacific Grove, Brooks Cole. 2000.

179 STEFANAKIS, Emmanuel, VAZIRGIANNIS, Michael, SELLIS, Timos. *Spatial decision making based on fuzzy set methodologies*. In : *XVIII INTERNATIONAL CONGRESS FOR PHOTOGRAMMETRY AND REMOTE SENSING*. Vol. XXXI, *Part*

B4, 1996, Viena. **Anais...** Viena : Karl Kraus, Diretor do Congresso, 1996. p. 829-834.

1180 STAR, Jeffrey, ESTES, John. ***Geographic information systems: an introduction***. Nova Jérsei : Prentice-Hall, Inc. 1990. 303p.

1181 TABALIPA, A. R. **Bancos de dados convencionais e o conceito de *prevalência***. Brasília, Faculdade Alvorada, 23 jul. 2002. Informação verbal.

1182 TANENBAUM, Andrew S. **Organização estruturada de computadores**. Rio de Janeiro: LTC S.A., 1990. 460p.

1183 THOMÉ, Rogério. **Interoperabilidade em *geoprocessamento* : conversão entre Modelos Conceituais de Sistemas de Informação Geográfica e Comparação com o Padrão *Open* GIS**. São José dos Campos, 1998. 200f. Dissertação (Mestrado em Computação Aplicada) - Instituto de Pesquisas Espaciais.

1184 TRIPODI, T., FELLIN, P., MEYER, H. **Análise da pesquisa social**. Rio de Janeiro : Livraria Francisco Alves Ed. 1975. 338p.

1185 TUFFANI, M. **Tradução traidora**. Capturado em 06 ago. 2002. *Online*. Disponível na Internet http://www.uol.com.br/cultvox/novos_artigos/critica_traducao. pdf

1186 TVERSKY, Amos. *Features of Similarity*. ***Psychological Review***, Jerusalém, v. 84, n. 4, p. 327-352. 1977.

1187 URDANETA, Iraset Paez. ***Gestión de la inteligência: aprendizaje tecnológico y modernización del trabajo informacional; retos y oportunidades***. Caracas : Universidade Simon Bolívar. 1992. 253.

1188 UNESCO. **A nova história da cartografia ou a história de uma nova cartografia?** Capturado em 12 ago. 2001. *Online*. Disponível na Internet : http: //www2. prudente. unesp. br/cartosig/Cartografia/Historia/historia. Html

1189 UNIVERSIDADE DA CAROLINA DO NORTE (UCN). ***Factor analysis***. . Capturado em 24 jul. 2002. *Online*. Disponível na Internet http: //www2. chass. ncsu. edu/garson/pa765/factor. Htm

1190 USCHOLD, M. ***Building ontologies : towards a unified methodology***. Edinburgh : University of Edinburgh. 1996. 18p.

1191 VICKERY, B. C. ***Classification and indexing in science***. London : Butterworth & Co. (Publishers) Ltd. 1975. 274p.

1192 VIEIRA, P. R.. **O uso de dados de RADAR no estudo de situação do comandante**. Rio de Janeiro, 1998. 99f. Monografia (Doutorado em Aplicações Militares) - Escola de Comando e Estado-Maior do Exército.

1193 WILLIAMS, J. G., SOCHATS, K. M., MORSE, E. *Visualization.* ***Annual rfeview of Information Science and Technology*** (ARIST), Medford, v. 30, p.161-207. 1995.

1194 WONG, Wai-Chiu. ***Incremental document clustering for web page classification.*** Capturado em 05 mar. 2002. *Online*. Disponível na Internet : http: //citeseer. nj. nec. com/328087. html

1195 WONNACOTT, Thomas H., WONNACOTT, Robert J. **Introdução à estatística**. São Paulo : LTC Ed. 1980. 589p.

1196 WUESTEFELD, Klaus. (klaus@objective.com.br) **Você ainda usa banco de dados? um questionamento céptico**. 06 abril. 2003a. Enviada às 05h. 09 min. Mensagem para: Paulo César Rodrigues Borges (pcrb@terra.com.br)

1197 WUESTEFELD, Klaus. ***Prevayler: breakthrough in memory technology***. Capturado em 05 abril. 2003b. *Online*. Disponível na Internet : http:// www. prevayler. org/ wiki. jsp? topic=BreakthroughsInMemoryTechnology

1198 YONG, Chu Shao. **Banco de dados : organização, sistemas e administração**. São Paulo : Ed. Atlas S.A. 1983. 398p.

APÊNDICE - A

Questionário formulado ao pessoal do quadro técnico da DSG e do CCAuEx

QUESTIONÁRIO

Parte 1: Instruções ao Respondente

1. Este questionário tem a finalidade de servir como subsídio para uma pesquisa sobre o problema da interoperabilidade entre sistemas de informações geográficas (SIGs) e sobre meios de facilitar a interação de usuários não especializados com estes sistemas.
2. Preencha todas as lacunas das perguntas do questionário. Há lacunas na **Parte 2** (Informações Gerais do Respondente) e seis perguntas na **Parte 4** (Perguntas).
3. Os espaços destinados para identificação e para o posto ou graduação do respondente na Parte 2 não são de preenchimento obrigatório. O respondente fica à vontade para preenchê-los ou não.
4. Antes de responder à Parte 4, o respondente deve ler as definições da **Parte 3** (Definições das Entidades Geográficas). As respostas da Parte 4 devem ser dadas única e exclusivamente em função dessas definições. O respondente deve tentar não associar mentalmente as entidades geográficas às suas formas gráficas (desenho) e responder ao que é pedido com base nas definições dadas e, no máximo, complementando essas definições com a sua experiência de vida em contacto com essas entidades geográficas.
5. Em cada pergunta da Parte 4, há uma entidade geográfica (natural ou artificial) que deve ser comparada com dez outras, logo a seguir. Nos parênteses ao lado de cada uma dessas dez entidades, o respondente deve julgar a sua semelhança com a entidade que é dada no cabeçalho da pergunta. A ordem de semelhança é a seguinte: "1" para a mais semelhante e "10" para a menos semelhante.
6. Ao terminar, entregue o questionário ao aplicador.

Obrigado!

Paulo César Rodrigues Borges – Ten-Cel QEM/QEMA
Doutorando da UnB em Ciência da Informação

Parte 2: Informações Gerais do Respondente

Marque com um "X" ou preencha as lacunas a seguir:

- **Nome (opcional):** _____
- **Posto/graduação (opcional):** _____
- **Idade:** _____ anos.
- **Local de nascimento:** _____ (UF)
- **Residência atual:** _____ (UF)
- **Sexo:** ( ) masculino ( ) feminino.

- **Participei diretamente do Projeto de Modelagem do Espaço Geográfico Brasileiro da 1ª DL:** ( ) Sim
- **Nível de escolaridade:**

( ) 2º grau completo

( ) 3º grau (graduação ou bacharelado universitário) incompleto

( ) 3º grau (graduação ou bacharelado universitário) completo

( ) 4º grau (pós-graduação) completo ou não

- **Idiomas (além do português):**

	LÊ	ESCREVE	FALA
Inglês			
Espanhol			
Francês			
Alemão			
Outro (Qual? _____)			

Parte 3: Definições das Entidades Geográficas

Aeroporto: extensa área (normalmente pública), situada em locais apropriados de uma cidade, em que se instalam edificações e outras estruturas adaptadas para o pouso e decolagem de aeronaves dos mais diversos portes. Essas estruturas também são capazes de manter mercadorias em depósito temporário, permitir o fluxo e a acomodação de pessoas e de garantir manutenção e abastecimento das aeronaves que delas se utilizem.

Armazém: edificação pública ou privada, normalmente situada num pátio especializado para algum tipo de transporte (hidroviário, ferroviário, rodoviário ou aeroviário), com a finalidade de armazenar mercadorias por um período variável de tempo.

Atracadouro: local de um cais ou de um porto (marítimo, lacustre ou fluvial), dotado de estruturas capazes de estabilizar e fixar embarcações para carga e descarga de pessoal ou mercadorias.

Bosque: grande arvoredo, não tão fechado como uma floresta, cujas árvores podem ter se originado de forma natural ou pela intervenção humana (plantada).

Cachoeira: seção do curso de um rio, causada pela resistência à erosão oferecida por rochas muito duras. Conforme a resistência dessas rochas tende a diminuir, a cachoeira transforma-se numa corredeira.

Caminho carroçável: faixa de terreno por onde se pode transitar a pé ou em carroças de tração animal.

Cais: elevação de terra, ordinariamente lajeada e murada, à beira de um rio ou de um porto, destinada ao embarque e desembarque de pessoas ou mercadorias.

Campo de futebol: área normalmente situada ao ar-livre ou num estádio, destinada à prática do futebol.

Campo de tiro ao alvo: área normalmente situada ao ar-livre ou num local apropriado de uma praça de esportes, dotada de pessoal e estrutura capazes de dar suporte à prática desportiva do tiro ao alvo.

***Campus* universitário:** extensa área (em geral pública) destinada a acolher os meios necessários em pessoal, edificações e equipamentos que garantam o desenvolvimento de atividades acadêmicas de ensino e pesquisa em diversas áreas do conhecimento.

Canal: curso d'água, projetado e cavado pelo homem, para suprir, em geral, a falta de rios ou riachos navegáveis, numa dada região, ou para distribuir águas captadas numa planície para a cultura.

Estação ferroviária: edificação situada ao longo de uma linha férrea, dotada de estruturas para carga e descarga de material, receber pessoal e redirecionar os comboios.

Estádio: ampla área de lazer, construída normalmente sem telhado ou cobertura, capaz de abrigar campo de futebol, pista de atletismo e outras estruturas que dêem suporte a estas modalidades desportivas e aos seus espectadores.

Estrada de ferro: via de comunicação destinada ao transporte de cargas e de pessoal. Os meios de transporte para esta modalidade de comunicação terrestre são máquinas e vagões, sustentados por trilhos.

Lago: grande extensão d'água, geralmente doce, inteiramente rodeada de terra.

Morro: forma isolada de elevação do terreno que não ultrapassa cerca de 300 m.

Museu: edificação pública que contém acervos documentais e culturais diversos, destinada à visita de estudantes ou turistas, interessados em pesquisa ou entretenimento.

Olho d'água: nascente ou fonte d'água natural perene que aflora no solo.

Pântano: trecho de terras inundadas, de extensão variável, normalmente impróprio ao cultivo.

Pátio ferroviário: extensa área situada ao longo de uma estrada de ferro, em que se instalam edificações e outras estruturas adaptadas para o recolhimento de máquinas e vagões de transporte ferroviário, para manutenção ou controle.

Piscina: fosso construído para ser preenchido com água tratada, destinado à prática desportiva da natação.

Praça de esportes: logradouro público em que se instalam diversas edificações (estádios, ginásios) e espaços (campos e arquibancadas), apropriados à prática desportiva.

Recife: uma ou mais formações rochosas à flor da água (normalmente salgada).

Represa: obstáculo construído transversalmente ao curso de um rio, riacho ou canal, com o fim de aproveitar a água retida para acionar turbinas de produção elétrica.

Riacho: rio pequeno ou regato, impróprio para a navegação.

Rio: curso d'água de grande extensão e de vazão variável, podendo ser total ou parcialmente navegável.

Vala: escavação ou fosso de largura limitada e profundidade variável, aberta, em geral, para liberar a terra de um excesso d'água, para irrigar um terreno árido ou mesmo para demarcar um limite de propriedade.

Vila: é um aglomerado esparso de habitações individuais ou coletivas, sem autonomia administrativa, dotado dos recursos de saneamento básico normalmente existentes num município.

Parte 4: Perguntas

Pergunta nº 1: Qual é a ordem de semelhança entre RIO e as dez entidades geográficas seguintes ? (lembre-se: 1 para a mais similar e 10 para a menos similar):

( ) Lago
( ) Bosque
( ) Vala
( ) Olho d'água
( ) Caminho carroçável
( ) Estrada de ferro
( ) Canal
( ) Cachoeira
( ) Pântano
( ) Riacho

Pergunta nº 2: Qual é a ordem de semelhança entre LAGO e as dez entidades geográficas seguintes ? (lembre-se: 1 para a mais similar e 10 para a menos similar):

( ) Rio
( ) Pântano
( ) Canal
( ) Cachoeira
( ) Olho d'água
( ) Estrada de ferro
( ) Caminho carroçável
( ) Bosque
( ) Riacho
( ) Vala

Pergunta nº 3: Qual é a ordem de semelhança entre ÁREA DE LAZER e as dez entidades geográficas seguintes ? (lembre-se: 1 para a mais similar e 10 para a menos similar):

( ) *Campus* universitário

( ) Represa
( ) Campo de futebol
( ) Praça de esportes
( ) Campo de tiro ao alvo
( ) Estádio
( ) Museu
( ) Estação ferroviária
( ) Vila
( ) Piscina

Pergunta nº 4: Qual é a ordem de semelhança entre CAIS e as dez entidades geográficas seguintes ? (lembre-se: 1 para a mais similar e 10 para a menos similar):

( ) Aeroporto
( ) Pântano
( ) Armazém
( ) Atracadouro
( ) Recife
( ) Estrada de ferro
( ) Lago
( ) Rio
( ) Pátio ferroviário
( ) Morro

Pergunta nº 5: Qual é a ordem de semelhança entre CANAL e as dez entidades geográficas seguintes ? (lembre-se: 1 para a mais similar e 10 para a menos similar):

( ) Riacho
( ) Vala
( ) Bosque
( ) Olho d'água
( ) Caminho carroçável
( ) Estrada de ferro
( ) Cachoeira
( ) Canal
( ) Pântano
( ) Lago

Pergunta nº 6: Qual é a ordem de semelhança entre ESTÁDIO e as dez entidades geográficas seguintes ? (lembre-se: 1 para a mais similar e 10 para a menos similar):

( ) Represa
( ) *Campus* universitário
( ) Piscina
( ) Praça de esportes
( ) Campo de tiro ao alvo
( ) Museu
( ) Estádio
( ) Estação ferroviária
( ) Vila
( ) Campo de futebol

Entregue o questionário ao aplicador.

APÊNCIE - B

Instruções para montar o ambiente para o PRONTO e outros documentos

O CD preso à contracapa final do tomo contém:

- O arquivo leia_me.txt, que explana a organização do material suplementar à tese e fornece algumas orientações para fazer funcionar o PRONTO©;
- Glossário (V. Cap. 11);
- Diagramas (de classe e de caso de uso);
- Tabelas de consolidação das respostas às seis perguntas do questionário;
- Ontologia de trabalho (Faxinal9a.xml);
- Código-fonte para execução do PRONTO©.

10. DIREITOS AUTORAIS E PERMISSÕES DE USO

Os produtos mais comuns como o sistema operacional (*Windows XP® Home Edition*, v. 5.1) e o pacote de automação de escritório da *Microsoft Office*, particularmente o *Word®*, o *Excel®* e o *PowerPoint®* não foram citados extensivamente porque foram adquiridos juntamente na compra do *notebook* do pesquisador (*Toshiba Satellite* 1805-S204).

Os produtos listados a seguir foram os mais citados ou utilizados diretamente no âmbito desta pesquisa.

MC da folha Faxinal (PR)©. Autoria da DSG. Cessão ao pesquisador, nos domínios estritos das necessidades de pesquisa desta tese de doutorado, com o alcance cronológico de utilização limitado ao término da defesa de tese.

Gothic®. Fabricado pela *Laser-Scan*, Inc., utilizado em consonância com a cobertura contratual de compra e venda de licença entre o fabricante e a DSG.

Obs: o recurso ***on the fly©*** do sistema Gothic® é uma denominação de autoria do núcleo de desenvolvimento da *Laser-Scan* para o módulo de atualização automática de topologia das bases de dados espaciais deste sistema.

Java ®. Fabricado pela *Sun Microsystems*, Inc., na forma do pacote *Java* 2 SDK (*standard edition*), versão 1.2.1, em um CD-ROM, para a plataforma *Windows*[312] 9x, adquirido na compra do livro de DEITEL (2000).

Java Creator ®. Fabricado pela *Xinox Software*, (versão 2.50 LE – *freeware* -, capturado em www.jcreator.com).

PRONTO©. Concepção e requisitos de Paulo Cesar Rodrigues Borges (crborges@terra.com.br). Projeto e implementação de José Inácio Leiria (zeleiria@terra.com.br).

SPSS™. Produzido pela *SPSS* Inc. (www.spss.com). Utilizado segundo as condições de licença estabelecidas na aquisição do aplicativo (SPSS™ v. 10.0) pelo CID/UnB.

TBCD (Tabelas da Base Cartográfica Digital)©. Produzidas pela DSG. Utilizada com autorização do Diretor do Serviço Geográfico do Exército Brasileiro.

---

[312] Windows™: Microsoft Corp.

11. GLOSSÁRIO DE TERMOS DAS GEOCIÊNCIAS E DAS CIÊNCIAS COGNITIVAS

Com o intuito de ampliar o conhecimento do leitor neste trabalho, que explora conceitos das Ciências da Terra e das ciências cognitivas, serão apresentadas definições destas áreas, que não são essenciais para o entendimento da metodologia desenvolvida na pesquisa. O conteúdo do subitem 1.5 (definições preliminares) já delineou o contexto no qual se insere esta pesquisa, de certa forma complementando a revisão de literatura. Este glossário, por conseguinte, é um suplemento.

Para não avolumar ainda mais o tomo principal do trabalho, o caderno suplementar que contém o glossário segue no CD apenso à capa de fecho do trabalho (Apêndice B), com os termos abaixo tabelados:

Abstração	**Generalização**	**Propriedades**
Aerotriangulação (fototriangulação)	**Inferência**	**Psicofisiologia**
	Inteligência artificial (IA)	***Reambulação***
Algoritmo genético (AG)	**Interdisciplinaridade**	**Reengenharia**
CAD	**Lógicas**	**Relevância**
Características	**Linguagem de representação do conhecimento (LRC)**	**Representação do conhecimento**
Cartografia		
CEPAD	**Métodos de modelagem orientada a objetos (MMOO)**	**Resolução**
Conceito inadequado		**Restituição**
Dado cartográfico	**Multidisciplinaridade**	**Sensores remotos**
Dado matricial	***Navalha de Ockham***	**SIG**
Dado Vetorial	**Neurocomputação**	**Sistema nervosos central**
Detecção Remota	**Neurofisiologia**	**TBCD©**
Digitalização	**Neuropsicologia cognitiva**	**Teoria do conhecimento**
DTD	**OGC**	***Tesauro***
Engenharia reversa	***Pixel***	**Topologia**
Fotogrametria	**Porta lógica**	**Transdisciplinaridade**
		User-friendly