2
Um ponto de vista fértil

Esta tese trata da elaboração automática de ontologias, inserindo-se na linha de pesquisa de processamento de linguagem natural (PLN). Uma vez que ontologias dizem respeito à descrição do mundo (ou de porções dele), e que "o projeto de dizer o que uma coisa é coincide inescapavelmente com a tentativa de circunscrever o significado de uma expressão lingüística" (Martins, 1999:137), a tese trata também, ainda que tangencialmente, de questões relacionadas ao significado, aproximando-se então do terreno movediço da semântica.

A questão "o que é o significado de uma palavra" é um dos problemas nucleares da investigação semântica. De forma bastante simplificada, é possível distinguir três paradigmas que irão problematizar o significado de forma sistemática: realista, mentalista e pragmática. Porém, ainda que didaticamente esta distinção seja útil, na prática, teorias realistas e mentalistas têm historicamente compartilhado pressupostos teóricos fundamentais, o que permite, com alguma simplificação, agrupá-las sob o rótulo *representacionistas* ou *essencialistas* (Martins, 2004).

Em uma visão mentalista, as palavras possuem uma relação estável com entidades mentais, isto é, a um significado corresponde um conceito, uma idéia. Já em uma visão realista, as palavras possuem uma relação estável com a realidade, com entidades do mundo que, por sua vez, podem ser reais ou virtuais. Para ambos, a linguagem é um sistema de representações de significados fixos e compartilhados; palavras representam *algo* (entidades mentais para os primeiros e virtuais para os segundos), e essa relação de representação se dá de maneira objetiva e estável.

Já o ponto de vista pragmático diz respeito à linguagem em uso, em diferentes contextos, considerando o uso feito pelos falantes na comunicação – o foco está na linguagem enquanto forma de *interação social*. O significado é decorrência de situações concretas, variáveis. Há, portanto, uma mudança de perspectiva, já que a linguagem passa a ser entendida como uma prática

intersubjetiva. Dentre as linhas de investigação pragmáticas, contudo, há as que poderiam ser também enquadradas em um paradigma essencialista. Isto porque se, por um lado, mentalistas irão assumir que é pela análise das propriedades dos códigos de linguagem que será possível explicar a prática da comunicação, algumas correntes da pragmática recomendam a análise das propriedades da prática da comunicação como maneira de fornecer uma explicação do que são as línguas e os significados, o que faz com que esta visão pragmática tradicional possa ser compreendida como uma disciplina complementar a uma visão semântica essencialista (Martins, 1999; Taylor, 1992). Porém, a crítica que pragmatistas mais radicais farão é que qualquer análise essencialista da linguagem é impossível, por ser impossível um distanciamento do objeto examinado; há uma relação mútua indissociável – nossas práticas de vida constituem a linguagem e, ao mesmo tempo, são por ela constituídas, o que impossibilita a realização de julgamentos absolutos sobre a linguagem. A relação entre linguagem e realidade seria forjada, na medida em que a própria linguagem constitui a realidade:

> "o que está sobrando é a pergunta 'Como a linguagem se liga à realidade?'. Pois se baseia firmemente em uma linguagem equivocada."
>
> (Hacker e Backer, 1984a :135)

Assumo neste trabalho uma postura compatível com uma visão pragmática radical do significado, segundo a qual a dificuldade em se responder à pergunta *o que é o significado* se deve à natureza equivocada da pergunta. A linguagem diz não o real em si, mas as opiniões e práticas dos homens, e por isso sua imprevisibilidade não é um desvio, mas conseqüência dessas opiniões ou impressões, que são naturalmente contraditórias (Martins, 2004).

Para lidar com o significado no ambiente do PLN, me apóio principalmente no ângulo sugerido por Wittgenstein, sobretudo nas *Investigações Filosóficas* (1953). É importante salientar, contudo, que Wittgenstein não apresenta uma teoria semântica, uma teoria do significado, ou mesmo uma filosofia da linguagem. Uma de suas grandes preocupações é mostrar que a linguagem não é um fenômeno único, e se furta a generalizações e sistematizações; o que ele propõe é uma elucidação do significado das palavras por meio da descrição de seu uso. Ao apontar para a resistência da linguagem à

investigação científica, Wittgenstein parece tematizar especificamente a questão do sentido na linguagem, sugerindo a inadequação da busca por uma ciência do significado (Martins, 1999).

Porém, assumir a inadequação da questão *o que é significado* não significa a defesa de uma posição reducionista segundo a qual significados não existem. Eles existem, mas não como entidades autônomas, e não com a precisão ou os limites definidos, necessários à formalização que sempre se buscou fazer. O significado é flexível e maleável, não cabe no molde fixo que lhe desejam impor. E, se esta recusa dos significados a uma formalização exaustiva pode ser uma forte limitação para as semânticas formais, por outro lado, pode representar uma motivação para outras formas de lidar com o significado. O significado não é uma propriedade imanente à palavra, mas uma função que expressões lingüísticas exercem em um contexto específico e com objetivos específicos (Marcondes, 2005). Com isso, o significado de uma palavra pode variar conforme o contexto em que é utilizado, conforme o objetivo desse uso.

Se não há uma essência única e fixa do significado, como lidar com as definições? Dicionários não só existem como são úteis. Negar esse fato parece um contra-senso. Porém, o que Wittgenstein enfatiza é o caráter parcial e incompleto das definições – que nem por isso as torna menos úteis. Desse modo, se, em uma perspectiva essencialista, esbarraríamos, em algum momento, nos "indefiníveis" – isto é, traços ou universais como "humano" ou "masculino", que compreendemos sem dificuldade – Wittgenstein argumenta que as definições são sempre fundamentadas em um conhecimento prévio, derivado do uso (do contexto, da situação de explicação, de inúmeros outros fatores). Isto é, definições, embora úteis nos contextos em que são utilizadas, serão sempre parciais. Explicações são sempre correlatas a pedidos de explicação, de modo que o significado é explicitado principalmente em situações que buscam desfazer mal-entendidos:

> Isso será feito (a descrição do uso de uma palavra, dizendo que objeto essa palavra designa) quando se tratar apenas de desfazer o mal entendido seguinte: pensar que a palavra lajota se relacione com a forma da pedra de construção que nós de fato nomeamos "cubos"' – mas o modo dessa 'relação', isto é, o uso dessas palavras, no restante, é conhecido."
>
> (Investigações Filosóficas, § 10)

Definições analíticas, que analisam termos com base em uma conjunção de marcas características, deixam de ser encaradas como "as" definições por excelência: trata-se apenas de mais uma forma de explicação, dentre outras possíveis. E, justamente por ser dependente do uso, dependente de uma situação concreta, e não uma entidade autônoma, a descrição do significado de um termo dificilmente se adequará ao formato das definições analíticas, composicionais e essenciais. Tais estratégias serão sempre limitadas:

> E o que ocorre com a última elucidação dessa cadeia? (Não diga "Não há nenhuma 'última' elucidação". É exatamente o mesmo que dizer: "Não há nenhuma última casa nesta rua ; pode-se sempre construir mais uma".)
>
> (Investigações Filosóficas, § 29)

É importante frisar que Wittgenstein não nega a validade de definições analíticas – definições analíticas são apenas um dos tipos possíveis de explicação, e enquanto tais são lances legítimos no *jogo de linguagem*[2] - , apenas lembra que elas são parciais, e não fundacionais na linguagem. Isto porque é impossível tomar distanciamento no jogo, isto é, parar de jogar e observá-lo de um ponto de vista exterior. Podemos fornecer explicações, generalizações, mas tudo isso consiste, ou já está previsto, no próprio jogo. Explicações, portanto, enquanto lances no jogo, funcionarão, isto é, servirão aos objetivos pretendidos, quando aplicadas às situações em que são produzidas, e não em todas as situações possíveis. Por isso, não são exaustivas, não são completas em si mesmas, não são absolutas (Martins, 2000).

Nesse sentido, a incompletude inerente às definições é uma faceta da ausência de um ideal de exatidão. Precisão e exatidão, novamente, são relativos. Não há um padrão único que as governe; a precisão é uma questão de adequação às circunstâncias e aos propósitos.

---

[2] O termo *jogo de linguagem* é utilizado de diferentes maneiras, em diferentes situações, sem, contudo, jamais ser explicitamente definido. Como observa Perloff, o termo é "difícil de entender, de vez que Wittgenstein, como é típico, jamais o define de forma plena, optando, em vez disso, por usá-lo freqüentemente, de um modo que ele acaba por tornar-se nosso" (Perloff, 1996:60, apud Martins, 1999:154). Detenho-me aqui no uso da expressão enquanto forma de "enfocar mais de perto as nossas atividades lingüísticas reais, descrevendo-as contra o pano de fundo de nossas práticas não lingüísticas" (Glock, 1997: 228). Fazem parte dos jogos de linguagem *atos de fala*; atividades mais complexas como *contar histórias*, *formar hipóteses e*

> "É inexato se eu não indicar a distância que nos separa até o sol até exatamente 1 m? E se eu não indicar ao marceneiro a largura da mesa até 0,001 mm?
> **Um** ideal de exatidão não está previsto; não sabemos o que devemos nos representar com isso – a menos que você mesmo estabeleça o que deve ser assim chamado. Mas ser lhe á difícil encontrar tal determinação; uma que o satisfaça."
> (Investigações Filosóficas, § 88)

A língua é naturalmente vaga, imprevisível e ambígua, e grande parte de sua robustez se deve justamente a isso. Nem todos os conceitos, porém, são realmente vagos, e, embora a maior parte dos conceitos empíricos admita casos fronteiriços, nem por isso se tornam inúteis (Glock, 1997).

> "It is precisely the lack of clarity in our use of the word culture which makes it such a handy word to have at one's disposal."
> (Stock, 1983 apud Kilgarriff, 1997: 39)

Na transposição das idéias de Wittgenstein para a lingüística, sigo aqui o caminho apresentado por Helena Martins (1999), segundo o qual uma lingüística descritiva compatível com o espírito wittgensteiniano ambiciona

> "fornecer descrições parciais e contingentes de regularidades observáveis nas línguas do mundo – sem pretender dar conta dos jogos como um todo, a partir de algum ponto de vista exterior(...)"
> (Martins, 1999:144)

Uma lingüística que

> "ambicionará, em seu impulso genuína e legitimamente generalizador, manter-se nos limites da linguagem, apresentando não uma visão verdadeira e completa dos fatos, antes um ângulo fértil pelo qual se possam reconhecer regularidades em nossas práticas lingüísticas."
> (Martins, 1999:144)

A perspectiva de Wittgenstein, por assumir nossa imersão no jogo da linguagem, é capaz de acomodar um ecletismo, uma visão mais tolerante com relação às diferentes teorias de linguagem. Com isso possibilita o uso, por exemplo, de um vocabulário tradicional, compreendido como lance no jogo de linguagem – do jogo de falar sobre a linguagem. Conseqüentemente, embora adotando o ponto de vista wittgensteiniano, não me privo, em diversos momentos,

---

testá-las; modos de discurso como *falar sobre objetos físicos* e jogos de linguagem de *falar sobre a linguagem* (Glock, 1997; Martins 1999).

da utilização de um vocabulário tradicional – em especial, de termos como *sintagma nominal*, *palavras denotativas*, *vagueza* e *hiponímia* – embora estes devam ser compreendidos de maneira deflacionada. Como bem esclarece Martins:

> "qualquer teoria sobre as línguas naturais será uma descrição parcial e reificadora de práticas sociais humanas – e isso vale tanto para as produzidas segundo um ângulo estruturalista quanto para aquelas que adotam o ponto de vista contextualista. (...) Continua fazendo algum sentido dizer, afinal, que o sistema verbal do português divide-se em três conjugações, ou que algumas línguas favorecem a omissão de sujeito na frase (...). Sem explicitar de maneira direta relações entre o lingüístico e o contextual, essas generalizações obviamente lançam alguma luz sobre nossos jogos de linguagem; se não alcançam a meta de revelar regras apriorísticas definitivas, pelo menos constituem descrições de regularidades envolvidas em nossas práticas lingüísticas."
>
> (Martins, 1999:146-147)

Lembro, por fim, que a idéia de que *explicações são sempre correlatas a pedidos de explicação* deve ser entendida de maneira abrangente. Assim, assumo aqui que o meu "*pedido de explicação*" é uma *aplicação* – uma ontologia específica de domínio. As explicações oferecidas, portanto, não pretendem um esgotamento da questão, mas pretendem responder, de maneira satisfatória, ao pedido.

As relações capturadas entre as palavras[3] retratam descrições parciais e contingentes de modos como são usadas nos jogos de linguagem em uma determinada língua (Martins, 1999), e acrescento, no caso específico deste trabalho, em um determinado domínio. Categorizar também é um jogar um jogo de linguagem com palavras.

2.1. O tratamento do significado no Processamento de Linguagem Natural

Tradicionalmente, a semântica computacional está ancorada em visões essencialistas-representacionistas do significado. As wordnets (Fellbaum, 1998; Vossen, 1998), por exemplo, são bases de dados lexicais que contêm "nomes,

[3] A delimitação de unidades lexicais é um tema controverso na teoria lingüística (Basílio, 1999). Neste trabalho, uso indistintamente "palavra" e "termo" para fazer referência às unidades lexicais.

verbos, adjetivos e advérbios agrupados em conjuntos de sinônimos cognitivos, *cada um representando um conceito distinto*"[4] (grifo meu).

Porém, nem sempre a diferença de abordagens com relação ao significado é nítida na Inteligência Artificial. Modelos como redes semânticas e enquadres fazem uso de inserção de conhecimento enciclopédico por um lado – incorporando elementos da pragmática tradicional – , e do formalismo de definição por traços e primitivos semânticos, por outro – incorporando elementos de um paradigma representacionista.

Além disso, se, como afirma Martins, "as idéias de Wittgenstein não têm comparecido em teorias lingüísticas com muita freqüência" (1999:136), no PLN a situação não é muito diferente – o que de forma alguma chega a ser surpreendente, visto a posição de Wittgenstein de não oferecer qualquer teoria unificadora sobre a linguagem e, principalmente, sua concepção de linguagem enquanto prática de vida que dificulta, por motivos óbvios, a possibilidade de transposições bem-sucedidas.

Em geral, o ponto de vista wittgensteiniano irá influenciar abordagens estatísticas do significado, especificamente aquelas voltadas para as tarefas de desambigüização de itens lexicais, em grande parte devido ao slogan "o significado está no uso". Nestes casos, a aproximação se dá por meio da substituição de *uso* pelo corpus; especificamente, pelas adjacências de uma palavra. Em termos gerais, calcula-se, para uma dada palavra-alvo, o número de palavras que aparecem ao seu lado em uma janela de tamanho pré-determinado - por exemplo, 15 palavras. Em seguida, cada palavra é representada por meio das freqüências cumulativas das ocorrências no escopo da janela. Palavras com significados similares tenderão a ocorrer em contextos similares e palavras polissêmicas tenderão a ocorrer em contextos diferentes.

Porém, as aproximações entre este tipo de trabalho e uma posição relativista com relação ao significado devem ser vistas com alguma cautela. Freqüentemente, o slogan serve de fachada para um trabalho estatístico que opta pela praticidade da não-definição dos termos. Schütze (1998), que propõe um mecanismo automático para a discriminação de significados, e não para a desambiguação, justifica sua

[4] Disponível em http://wordnet.princeton.edu/. Acessado em 19/12/2006

escolha exatamente por ser a desambiguação dependente de uma definição do significado[5].

Widdows (2003) apresenta um modelo de aquisição e desambigüização lexical baseado em informação estatística contextual, no qual não existem definições de palavras, mas apenas relações entre os termos. Mas, embora dispense as definições, ele afirma que o significado deve poder ser descrito de forma "clara, flexível e acurada", através de um pensamento científico cuidadoso e de investigação empírica. Ainda segundo Widdows (2003), métodos estatísticos, embora tenham trazido enormes contribuições, apenas *imaginam* ou *supõem* o significado das palavras.

Os trabalhos de Adam Kilgarriff, claramente inspirados em posições relativistas do significado, são os mais afinados com a perspectiva teórica assumida aqui. No artigo "I dont believe in word senses" (1997), Kilgarriff, tomando o ponto de vista de um lexicógrafo, salienta que significados só existem com relação a um objetivo, que pode ser o de escrever dicionários ou tesauros, por exemplo. Lembrando ainda a escassa literatura sobre a utilização de critérios na tarefa subjetiva de separação dos significados das palavras – o que contribui para dificultar ainda mais o trabalho dos lexicógrafos –, Kilgarriff sustenta que uma lexicografia de corpus é a mais apropriada para o tratamento dos significados, uma vez que ofereceria uma resposta diferente para a questão do significado de uma palavra. Assumindo, com Wittgenstein, que as palavras só têm sentido no uso, o lexicógrafo deve recorrer ao corpus como se fosse ele, o lexicógrafo, um "instrumento" cuja função é organizar o que está no corpus e "traduzir" esta organização para a linguagem de definição de dicionário.

> "Word senses are simply undefined unless there is some underlying rationale for clustering, some context which classifies some distinctions as worth making and others as not worth making"
>
> (Kilgarriff, 1997: 107)

Em outro trabalho, voltando-se diretamente para o PLN (2003), Kilgarriff propõe que tesauros sejam construídos automaticamente a partir de corpora, com base não nos diferentes significados das palavras, mas nas próprias palavras. Por

[5] "Word sense discrimination is easier than full disambiguation since we need only determine which occurrences have the same meaning, and not what the meaning actually is"

meio da aplicação de algoritmos de agrupamento (*clustering*) sobre um corpus, o tesauro seria um agrupamento de termos relacionados – e o significado seria atribuído à palavra em função do grupo a que a palavra pertence.

Os trabalhos de Yorick Wilks também buscam aproximações com uma visão não-representacionista (Wilks, 1999; Niremburg e Wilks, 2001), e ultimamente Wilks tem se dedicado à investigação de ontologias (Wilks, 2002; Niremburg e Wilks, 2001; Brewster e Wilks, 2004).

Com relação à língua portuguesa, os trabalhos de Garrão (Garrão et al., 2006; Garrão, 2006), voltados para a identificação de expressões multi-vocabulares verbais, também incorporam um ponto de vista não-representacionista semelhante ao apresentado aqui.

Por fim, embora a enorme afinidade entre a perspectiva que assumo neste trabalho – de natureza predominantemente aplicada – e a crença de que esta perspectiva, principalmente nos termos de Martins (1999) é, de fato, promissora para estudos relativos ao significado, não acredito que esta afinidade seja condição necessária para o sucesso da investigação em PLN. Neste ponto, compreendo que um dos objetivos do PLN é a resolução de problemas. Assumo aqui, portanto, a perspectiva da IA fraca: basta que o desempenho dos programas imite o funcionamento da linguagem, não é preciso que os processos subjacentes, em ambos os casos, sejam os mesmos. Meu comprometimento, nesse sentido, é com resultados satisfatórios, e não com determinadas perspectivas teóricas, as quais são utilizadas na medida em que oferecem *insights* interessantes para a abordagem dos problemas. Concordo, portanto, com Diana Santos quando afirma que

> "(...) é ao tentar resolver um dado problema (isto é, ao tentar construir um programa que manipula a língua) que surge o momento de nos debruçarmos quer sobre (algumas características) do léxico ou da gramática, quer sobre as teorias que pretendam dar respostas a esse problema"
>
> (Santos, 2001:229)

A perspectiva wittgensteiniana de linguagem é compatível com essa visão, já que, para Wittgenstein, não é possível uma descrição completa da língua porque não é possível deixarmos de tomar parte no jogo para apenas observar; não é

(Schütze, 1998:97)

possível termos a visão do todo. Conseqüentemente, não há objeto ou processo a ser simulado. De fato, como afirma Sparck Jones,

> "The challenge of taking the necessary step from a focused experiment or even convincing prototype to a full-scale rounded-out NLP system with consistent, high-quality performance has not been overcome."
>
> (Sparck Jones, 2001:9)

Em conseqüência, acredito que a perspectiva teórica adotada para cada situação problema indica apenas que ela foi a mais produtiva para o tratamento daquele problema específico, mas não necessariamente que o será em outros casos. Enfim, para o tratamento de relações de significado entre as palavras, uma visão não-representacionista é um ponto de vista fértil.

2.2. Ontologias e significados – uma visão tradicional

O estudo das ontologias, embora desperte grande interesse no campo da Inteligência Artificial (IA), remonta às origens da filosofia, há cerca de 25 séculos. Mas esta longa tradição não significa que existam respostas satisfatórias aos problemas inicialmente apresentados. O termo, originalmente, designa o estudo do ser, considerado independentemente de suas determinações particulares e naquilo que constitui sua inteligibilidade própria. Trata-se da teoria do ser em geral, da essência do real (Japiassú e Marcondes, 1989). Enquanto teoria do ser, uma ontologia busca descrever as categorias mais básicas da realidade - entidades, tipos de entidades e o relacionamento entre esses elementos.

A investigação sobre as categorias que compõem a realidade começa a receber um tratamento sistematizado nas *Categorias*, de Aristóteles, que apresenta 10 categorias básicas que classificariam tudo o que pode ser dito ou predicado sobre qualquer coisa: substância, quantidade, qualidade, relação, lugar, tempo, posição, estado, atividade e passividade. O filósofo Franz Bretano, em 1862, adicionou alguns termos retirados de outros escritos de Aristóteles, e criou um diagrama de árvore como o da figura 1 (apud Sowa, 1999).

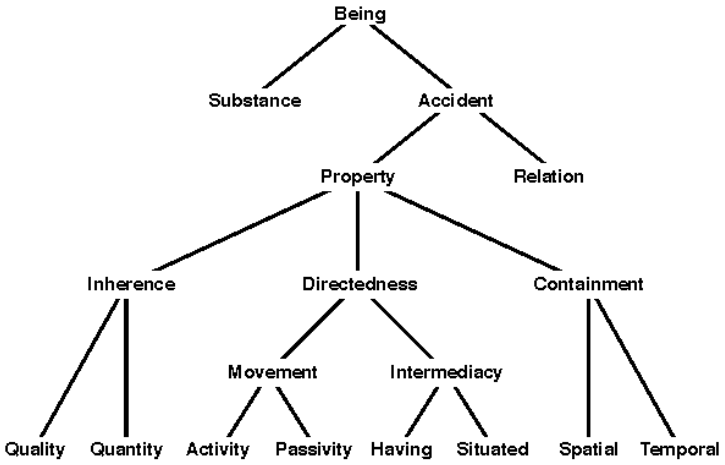

Figura 1: Categorias de Aristóteles, por Franz Bretano (apud Sowa, 1999)

As categorias expressas pela realidade descreveriam o real – assume-se a existência de um mundo externo à linguagem, passível de descrição. Ontologias devem, portanto, ser gerais e independentes de língua, pois descrevem a realidade, que, por sua vez, é a mesma para todos – e por isso os conceitos são gerais, independentes de língua. Ou seja, nessa visão, aos conceitos das ontologias são atribuídos rótulos – as palavras – , que serão dependentes de língua. De fato, essa é a perspectiva que norteia, até hoje, redes lexicais como as wordnets (Fellbaum, 1998; Vossen, 1998), que freqüentemente são utilizadas como ontologias:

> "In principle, the separation between ontology and lexicon is as follows: 'language-neutral' meanings are stored in the former; language-specific information in the latter."
>
> (Viegas et al., 1999: 21)

Na IA, a necessidade de formas padronizadas para a codificação do conhecimento foi reconhecida no início da década de 70. O ANSI (*American National Standards Institute*) propôs que todo o conhecimento pertinente sobre um domínio deveria estar concentrado em um único *esquema conceitual*, como ilustra a figura 2 (apud Sowa, 1999). A função de tal esquema seria fornecer definições comuns para as entidades das aplicações e explicitar o relacionamento entre elas (Sowa, 1999).

De acordo com Sowa (1999), por mais de 20 anos esse esquema conceitual foi importante no desenvolvimento e uso de aplicações integradas, mas nunca

houve implementações completas; nunca se atingiu o objetivo final de integração total em torno de um único esquema.

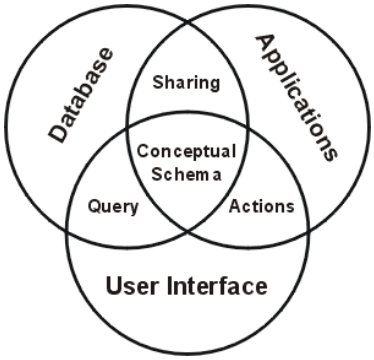

Figura 2: Esquema conceitual como núcleo de um sistema integrado (Sowa, 1999)

É nesse contexto que a IA se apropria do termo ontologia: o crescente reconhecimento de que fontes computacionais devem ser as mais gerais possíveis, reutilizáveis e compartilháveis entre a comunidade de IA foi o primeiro passo para considerar o valor das questões tradicionais da filosofia: o estudo da realidade e seus objetos, independentemente do nosso conhecimento sobre eles, e a busca por uma natureza a priori das coisas (Bateman, 1995). Para Guarino (1995), uma base de conhecimento que se aproximasse à noção filosófica clássica de verdade facilitaria não apenas a interação e comunicação entre diferentes agentes, mas também o compartilhamento e reaproveitamento da própria base.

Segundo Bateman (1995), no que tange às ontologias, há uma confluência apenas aparente de interesses entre filosofia e IA. Na IA, o uso do termo remeteria à construção de *frameworks* para "conhecimento" que permitam a sistemas computacionais lidar com problemas tais como processamento de linguagem natural e "*real world reasoning*". De acordo com essa perspectiva, um sistema deve ser capaz de realizar deduções com relação a algum corpo de informação, e os componentes organizacionais mais gerais desta informação são chamados coletivamente de ontologia. Guarino (1995) defende a introdução sistemática de princípios de ontologia formal na engenharia de conhecimento, a fim de explorar as várias relações entre ontologia e representação de conhecimento. Para a área de sistemas de informação, uma ontologia seria uma linguagem formal elaborada para representar um domínio particular de conhecimento, cujo objetivo é, essencialmente, funcional (Zúñiga, 2001). Em última análise, a própria discussão

sobre o que venha a ser uma ontologia é ilustrativa da dificuldade de se estabelecerem definições e conceitos comuns e compartilháveis entre domínios. Ou seja, a dificuldade em se chegar a um acordo sobre o que são ontologias põe em xeque a própria existência de ontologias nos moldes propostos – uma ontologia geral, multilingüe e, algumas vezes, independente de domínio.

De fato, a elaboração de ontologias sustentadas por representações de conhecimento gerais, independentes de língua, parece ser problemática. O projeto de construção de uma única ontologia, que pudesse ser ao mesmo tempo não-trivial e adaptável para diferentes comunidades de sistemas de informação, foi em grande parte abandonado; a tarefa se mostrou muito mais difícil do que o previsto inicialmente, confirmando os problemas já enfrentados por filósofos há 2000 anos (Smith, 2001).

O desapontamento com construção de ontologias gerais, levou, por sua vez, ao investimento em ontologias específicas de um domínio. Neste contexto, uma das definições de ontologia mais difundidas é a de Gruber (1993), segundo a qual uma ontologia é "uma especificação formal explícita de uma conceitualização compartilhada".

No âmbito da pesquisa em PLN, ontologias podem ser vistas como "modelos de domínios específicos", que têm como objetivo facilitar buscas semânticas (Brewster e Wilks, 2004).

2.3. Ontologias e significado – uma visão relativista[6]

Paralelamente à visão tradicional, desenvolve-se, na filosofia, uma outra abordagem, relativista, anti-essencialista, cujo embrião pode ser encontrado já no pensamento sofista, e que sustenta não existir uma realidade independente e exterior à linguagem e, portanto, passível de uma descrição essencialista. Segundo essa perspectiva pragmática radical, a própria empreitada ontológica perde o sentido – isto é, não se trata de uma tarefa difícil, mas de uma tarefa sem sentido: não há conceitos independentes de língua que descrevem o universo (ou parte dele) – em última análise, não há universo a ser descrito independente de língua, vista como práxis. O estabelecimento de verdades universalmente válidas,

[6] Refiro-me, no decorrer do trabalho, a um relativismo lingüístico-conceitual.

autônomas com relação às circunstâncias concretas é, do ponto de vista wittgensteiniano assumido neste trabalho, impossível. Somos constituídos pela linguagem, o que impossibilita a realização de julgamentos absolutos sobre ela. Ontologias gerais, aproximações às noções de verdade, não são questões que devam ser consideradas.

Mas, se não há "entidades mentais" ou realidade às quais as palavras se "colam", e que corresponderiam ao significado das palavras, o que é o significado então? A posição anti-essencialista de Wittgenstein, expressa principalmente nas *Investigações Filosóficas* (1953) e abordada no início deste capítulo, é de grande valia para lidar com o significado – intimamente relacionado à questão da elaboração de ontologias. Os significados correspondem aos usos culturalmente determinados que fazemos das palavras – o significado não é uma entidade, ele está no uso (Martins, 2004).

E no que as considerações de Wittgenstein podem ser úteis à semântica computacional, à elaboração de ontologias?

Na IA, como já mencionado, a ambição inicial de ontologias gerais foi substituída pela idéia de ontologias de domínio. Além da redução no escopo da tarefa, a constatação de que a elaboração de ontologias exige um processo longo de concordâncias entre um número grande de especialistas levou à pesquisa sobre formas de automação desse processo, considerando-se que o conhecimento a ser representado na ontologia deve ser a informação contida em textos (Buitelaar et al., 2005).

Adotando uma perspectiva relativista, na qual a linguagem e realidade se constituem mutuamente, é difícil pensar em ontologias baseadas em conceitos pré-definidos. Por outro lado, é igualmente difícil transpor a "linguagem enquanto prática de vida" para um ambiente computacional. Diante desse impasse, proponho a substituição (grosseira, é verdade) de "práticas de vida" pela informação contida no corpus – assumo que o conhecimento disponível em textos, expresso em linguagem natural, pode funcionar como uma fonte confiável para a busca de informações e categorizações.

Conseqüentemente, a principal característica da ontologia proposta é a ausência de categorias pré-definidas. Categorias em uma taxonomia são construtos humanos, abstrações que refletem uma perspectiva particular do mundo (Kilgarriff 2003, 1997; Brewster e Wilks, 2004). A idéia de sustentar a ontologia

em corpus busca deslocar o espaço de discussão sobre quais seriam as categorias relevantes de um domínio: do desejado consenso entre especialistas para as categorias motivadas pelo corpus, que, por sua vez, refletiriam o conhecimento implícito do domínio em questão.

2.4. Ontologias, tesauros e taxonomias

Como dito anteriormente, ontologias, no contexto de PLN, podem ser vistas como "modelos de domínios específicos", que têm como objetivo facilitar buscas semânticas. Neste ponto, surge outra confusão terminológica: no que diferem ontologias, tesauros, taxonomias e hierarquias? Trata-se de objetos cujas características se sobrepõem e que também compartilham, no PLN, do mesmo objetivo: auxiliar buscas semânticas. Em conseqüência, encontramos trabalhos muito parecidos mas que atribuem diferentes nomes aos seus "modelos": ora fala-se em tesauro (Kilgarriff, 2003), ora em ontologia (Vossen, 2003; Velardi et al., 2005; Brewster et al., 2005) e ora em taxonomia (Snow et al., 2005, Widdows, 2003).

Depois da discussão sobre o uso do termo ontologia apresentada na seção 2.2, volto ao tema, assumindo ontologia como caracterização de um domínio e explicitando diferenças e sobreposições com os outros termos.

Uma *taxonomia* é uma hierarquia de termos, na qual podem existir diferentes tipos de relação pai-filho (parte-todo; tipo-instância). Já um *tesauro* pode ser considerado uma extensão de taxonomia, comportando a inclusão de regras de uso de vocabulário, definições, sinônimos e antônimos. Compreende, portanto, além de relações hierárquicas, relações associativas. Por fim, *ontologias* (as específicas de domínio, pelo menos) são mais detalhadas; podem – e devem – conter mais níveis hierárquicos. Em um tesauro, as relações "termo genérico" / "termo específico" podem significar tanto uma relação de hiperonímia quanto de meronímia. A palavra *cachorro*, por exemplo, está relacionada a *mamífero*, em uma relação de hiperonímia; mas a palavra *cabelo* está relacionada a *corpo*, em uma relação de meronímia. Já os "termos associados" cobrem diversas relações semânticas, sem especificação. Alguns dos termos associados a cachorro são: *au-*

au, bassê, *labrador*, *latido*, *cadela*, *canil*, *cão* e *carrocinha*[7]. Nas ontologias, esta ambigüidade de relações não é possível: isto é, considerando, por exemplo, relações hierárquicas de meronímia e hiponímia, é preciso que haja uma distinção formal entre os dois tipos de relação – e não apenas um rótulo geral "termo geral – termo específico" que as abarque.

Neste trabalho entendo como ontologia um conjunto de termos, associados com definições em linguagem natural, que utiliza relações formais e é relativo a algum domínio de interesse (Hovy, 2002). Em termos gerais, trata-se de uma forma de organização do conhecimento de um domínio – o que também está de acordo com a definição de Gruber. Uma ontologia deve ser capaz de capturar uma série de relações semânticas entre termos, não apenas uma relação de inclusão de classe, como é o caso da relação de hiponímia. Ainda assim, assumindo que, em termos práticos, o resultado final deste trabalho é a elaboração automática de uma taxonomia (pois as relações extraídas são de hiponímia), uma vez que a metodologia adotada não impede a possibilidade de inserção de outros tipos de relação semântica, mantenho o uso de *ontologia*. Mas, reforçando o que foi dito, há consciência de que estou tratando, especificamente, de uma porção da ontologia – estou tratando de uma taxonomia.

2.5. Sobre taxonomias e hipônimos

Tradicionalmente, de um ponto de vista representacionista, é tarefa da semântica lexical representar o significado de cada palavra e explicar as relações sistemáticas entre esses significados (Saeed, 1997). Essas relações entre significados – nas quais a hiponímia se inclui – tomam por base, em termos gerais, a distinção clássica entre propriedades essenciais e acidentais. Especificamente, a estabilidade das relações é garantida pelas propriedades essenciais, imutáveis.

No caso da hiponímia, a inclusão de uma palavra numa classe feita com base em uma propriedade acidental não é considerada hiponímia. Isto é, uma relação como

[7] Exemplos retirados de Thesaurus da Língua Portuguesa do Brasil, disponível em http://alcor.concordia.ca/%7Evjorge/Thesaurus/. Acessado em 19/12/2006

gato < animal de estimação

não é uma relação de hiponímia válida porque toma por base uma propriedade como "*domesticável*", que não faria parte das propriedades essenciais de *gato*. Obviamente, esbarramos aqui na discussão sobre quais seriam as propriedades essenciais.

Esta seção descreve como se comporta a relação de hiponímia e sua conexão com a taxonomia de um ponto de vista tradicional, tomando como principais referências os trabalhos de David Cruse (1986, 2004) e John Lyons (1980).

Uma taxonomia pode ser considerada um tipo de configuração lexical, isto é, um vocabulário pode se estruturar em termos hierárquicos. Esse vocabulário tanto pode ser um vocabulário controlado, específico, que caracteriza as taxonomias científicas, quanto o vocabulário de uma língua natural, que por sua vez caracteriza as chamadas taxonomias populares (Lyons, 1980) ou naturais (Cruse, 1986).

Cruse (1986) e Lyons (1980) apontam uma série de características que distinguem as taxonomias formais (ou científicas) das naturais. Nas taxonomias formais, por exemplo, co-hipônimos devem ser incompatíveis: *melancia* e *abacaxi* são co-hipônimos e incompatíveis. Já nas línguas naturais, dois hipônimos do mesmo superordenado não são, necessariamente, incompatíveis: *romance* e *capa dura* (*hardcover*) são hipônimos de *livros*, mas não são incompatíveis – um *romance* pode ser de *capa dura*. De fato, este tipo de incompatibilidade (que é problema para Cruse (1986), mas não para Cruse (2004)) se deve justamente à imprecisão dos limites dos conceitos, o que será abordado mais adiante. Outra diferença entre taxonomias formais e naturais diz respeito à quantidade de níveis. As taxonomias naturais caracterizam-se por ter, no máximo, cinco níveis de profundidade, e mesmo esse número já é bastante raro. Já as taxonomias técnicas ou científicas não têm um número limitado de níveis. Por fim, considerando, de um ponto de vista formal, uma taxonomia ideal, todos os ramos possuem nós em cada nível. Quanto a isso, taxonomias naturais estão longe do ideal, pois estão repletas de lacunas lexicais (*lexical gaps*), isto é, termos para os quais não há um item superordenado. De fato, o vocabulário de uma língua natural pode estar

estruturado hierarquicamente a partir de diversos pontos de origem. Não há nenhum termo superordenado em relação a todos os outros, mas, segundo Lyons, "é inegável a existência de um certo grau de organização hierárquica em todos os níveis do vocabulário das línguas já estudadas" (1980:243).

Levando em conta as diferenças entre taxonomias naturais e científicas, a taxonomia apresentada neste trabalho é uma taxonomia híbrida: por um lado, apresenta características das taxonomias científicas – é específica de um domínio, e baseada em um corpus que contém textos técnicos. Por outro, as informações fonte para a sua elaboração vêm de textos - isto é, vêm de linguagem natural – o que a aproxima das taxonomias naturais.

A relação de hiponímia é a relação-chave de uma taxonomia. Trata-se de uma relação entre uma palavra mais específica (subordinada) e uma palavra mais geral (superordenada), como a relação entre *melancia* e *fruta*. Certamente a hiponímia pode ser considerada uma das formas mais importantes de estruturação do vocabulário, já que a inclusão dos termos em classes possibilita generalização, que se traduz em economia e aproveitamento de informação.

Do ponto de vista lógico, a hiponímia é caracterizada a partir de três critérios: (i) inclusão de classes; (ii) implicação unilateral; (iii) transitividade.

De acordo com o primeiro critério, tem-se que o item subordinado está incluído na classe do superordenado (ou, de modo inverso, que o item superordenado contém o item subordinado): *melancia* está incluída na classe das *frutas* (ou a classe das *frutas* inclui *melancia*).

O critério (ii), implicação unilateral, explora o fato de que a frase *Maria ganhou uma rosa* implica *Maria ganhou uma flor*, mas *Maria ganhou uma flor* não implica *Maria ganhou uma rosa*.

Por fim, o critério da transitividade, o que mais importa para a ontologia, pois permite a realização de inferências, mostra que hiponímia é uma relação transitiva: se X é hipônimo de Y e Y é hipônimo de Z, então X é hipônimo de Z: *melancia* é uma *fruta*; *fruta* é um *alimento*; então *melancia* é um *alimento*. Porém, como aponta Cruse (2004), há diversos casos em que a cadeia de transitividade parece se quebrar, principalmente se um dos elementos da cadeia não é um elemento prototípico: se um *banco de carro* é um *banco* e um *banco* é um *móvel*, então um *banco de carro* é um *móvel* não parece uma relação aceitável.

Além desses três critérios tradicionalmente utilizados, Cruse (1986) acrescenta os seguintes testes para a identificação da hiponímia:

a) equivalência a uma paráfrase em que o superordenado é modificado por um adjetivo:

Rainha é um monarca feminino

b) ocorrência em determinadas construções como:

Gatos *e outros* animais;

Não há flor *mais* bela *que* a rosa;

Ela gosta de *todas as* frutas, *exceto* manga;

Ela lê livros o dia todo – *principalmente* romances

Porém, logo em seguida, Cruse apresenta contra-exemplos que questionam os critérios propostos. Para o critério (a), como criar uma paráfrase equivalente a, por exemplo, *aranha*? A *aranha é um animal*... Já o critério (b) não oferece garantias de discriminação de hipônimos, pois consideraria como tais os elementos presentes em

(1) Gatos *e outros* animais de estimação;
(2) Cobras *e outras* criaturas venenosas;
(3) *Não há* arma *tão* versátil *quanto* uma faca.

Para Cruse (1986), nenhuma das expressões acima contém itens relacionados por meio de hiponímia, visto que as definições abaixo – que seriam fundamentais para a caracterização da hiponímia – não correspondem à realidade:

(1) ? um gato é necessariamente um animal de estimação
(2) ?uma cobra é necessariamente uma criatura venenosa
(3) ?uma faca é necessariamente uma arma

Como discorda de que os exemplos (1), (2) e (3) sejam exemplos de hiponímia, Cruse (1986) propõe uma subdivisão entre os hipônimos: os taxônimos (*taxonymys*). Taxônimos de um item lexical são um subconjunto de seus hipônimos, e seriam elementos cruciais para uma hierarquia lexical taxonômica – com isso, (1), (2) e (3) não seriam taxônimos, seriam apenas hipônimos. A diferenciação entre hiponímia e taxonímia poderia ser feita por meio do seguinte "contexto diagnóstico":

Um X é um tipo de Y

Se X é um taxônimo de Y, então o resultado é considerado normal:

(4) Um labrador é um tipo de cachorro.
(5) Uma rosa é um tipo de flor
(6) Uma banana é um tipo de fruta.

Nos exemplos acima, X é, de fato, um hipônimo de Y. Porém, nem todos os hipônimos levam a um resultado normal neste contexto:

(7) Uma rainha é um tipo de monarca
(8) Um garçom é um tipo de homem.

Segundo Cruse (1986), esses problemas decorreriam da multiplicidade de contextos em que a expressão "tipo de" é utilizada na linguagem cotidiana. Um dos contextos irrelevantes para a identificação ocorreria, por exemplo, em perguntas com o formato ambíguo como "*Que tipo de pessoa é ela?*" e "*Aquele tipo de pessoa que nunca paga suas contas*". Uma pergunta como "*Que tipo de árvore é aquela?*" provavelmente deseja uma resposta taxonômica. Por outro lado, alguém que pergunta "*Que tipo de árvore você está pensando em colocar no quintal?*" poderia muito bem se satisfazer com uma resposta "*Uma que dê bastante sombra*".

Ainda segundo Cruse (1986), "reconhecer uma taxonímia é uma coisa; *descrever sua essência natural* é uma outra tarefa, mais difícil" (1986:139) (grifo meu)[8]. Porém, ao invés de abandonar a distinção entre taxônimos e hipônimos, o autor segue tentando dissecar as diferenças entre as duas categorias, apoiando-se em algumas abordagens que, segundo ele, parecem iluminar a questão[9].

A primeira dessas abordagens diz respeito a uma forte correlação entre taxônimos e os chamados "tipos naturais" (*natural kind terms*), por um lado, e entre hipônimos não-taxonômicos e "tipos nominais" (*nominal kind terms*), por

[8] "Recognizing a taxonomy is one thing; describing its essential nature is another and more difficult task" (Cruse, 1986:139)

[9] "However, there are two or three lines of approach which seem to throw some light on the matter" (Cruse, 1986:139-140).

outro[10]. Porém, o próprio Cruse apresenta como contra-exemplo para essa distinção a taxonomia para "cores de cabelo" (hair-colour), uma taxonomia bem formada e que é baseada em tipos nominais[11].

A segunda consideração referente à caracterização da relação de taxonímia seria em termos de prototipicidade: categorias taxonômicas seriam constituídas por elementos prototípicos. Neste caso, poderíamos afirmar que a divisão taxonômica entre *garanhões* e *éguas* é taxonomicamente anômala, pois o critério *sexo* não seria o melhor critério para diferenciação de categorias. Mas Cruse também refuta esse argumento, afirmando que nem sempre a divisão em classes com base em prototipicidade é possível.

Cruse (1986) chega então à conclusão ("pessimista para a teoria semântica", segundo ele) de que

> "Perhaps (...) there are no invariable principles to be applied which inevitably lead to unique taxonomies; perhaps we merely seek to create the closest analogues we can to natural species. Exactly how close we get will of course depend on the nature of the category being sub-divided"
>
> (Cruse, 1986:144)

Finalmente, em Cruse (2004) a tentativa de distinção entre hipônimos e taxônimos é abandonada. Em nome do que chama "dynamic construal approach", o autor propõe um questionamento da assumida estabilidade do significado das palavras:

> "There are many different approaches to the study of semantic properties of words, but most of them take it for granted that each Word has a stable, inherent attribute called a 'meaning', which it is the job of a lexical semanticist to describe. (…) there is a general agreement that word meanings exist, and that logical and structural aspects of meaning, such as sense relations, and certain logical properties of utterances, are either directly represented in the lexical entry or can be inferred from the lexical entry.
> Of course, there is also general agreement that meaning is highly context-dependent (…).
> However, it has proved extremely difficult to achieve a satisfactory picture using these assumptions, and some linguists have begun to explore the consequences of abandoning the assumption of stable word meanings. (…)Well, the proposal is not

[10] Tipos naturais referem-se a classes de entidades que existem na natureza, como árvores, gatos, banana; e tipos nominais referem-se a agrupamentos mais arbitrários.

[11] Exemplo da taxonomia:
hair-colour>blonde, red-head, brunette;
blonde>ash blonde, strawberry-blonde

> that words have no stable semantic properties, but rather that these properties are not meanings"
>
> (Cruse, 2004: 261-262)

Nesta nova abordagem, palavras não têm significados que lhes são permanentemente atribuídos. Os significados emergem do uso como o resultado de vários processos interpretativos. O que as palavras possuem como uma propriedade permanente é um mapeamento entre um corpo de conteúdos conceituais (que Cruse chama de "*purport*"), que é parte essencial da matéria prima necessária ao processo de construção da interpretação (*construal process*), mas que sub-determina quaisquer significados específicos. Esses processos de interpretação que resultam em significados contextualizados estão sujeitos a restrições de vários tipos e com diferentes forças, o que torna determinadas leituras mais plausíveis que outras.

Esta nova abordagem de Cruse parece mais compatível com o enquadre wittgensteiniano assumido aqui:

> "it is an essential feature of the dynamic construal approach that, just as words are not associated with specific meanings, nor are they associated with specific conceptual categories, but with bodies of purport which allow variable construal in different contexts."
>
> (Cruse, 2004:267)

Assim, do mesmo modo que não podemos concordar com a abordagem de Cruse (1986) quando desconsidera relações de hiponímia como *gatos são animais de estimação*, concordamos com Cruse (2004) quando afirma que

> "Taking the dynamic construal view (...): *cat* is a hyponym of *pet* in *cats and other pets*, but not in *It's a cat, therefore it is a pet*, and the difference is due to the fact that the construed categories are different in the two contexts."
>
> (Cruse, 2004:268)

Porém, se tanto para a abordagem pragmática wittgensteiniana quanto para a abordagem de Cruse (2004) não há dificuldade quanto à aceitação de determinados termos em determinadas categorias, desde que possíveis em um contexto de uso, e ainda que a inclusão na classe seja pouco convencional, tanta "permissividade", mesmo que teoricamente motivada, leva a um problema significativo no âmbito do PLN: como então proceder a uma avaliação dos

resultados, isto é, como será possível saber se a identificação automática de relações de hiponímia é realmente eficaz, se o processo de inferências produziu conhecimento correto e como comparar esses resultados com o de outros trabalhos semelhantes? O capítulo 3 problematiza a questão da avaliação neste tipo de pesquisa.