2
Influências

Este capítulo aborda as principais influências teóricas para o desenvolvimento do modelo que será apresentado.

Nosso trabalho se apóia especialmente nas obras *A gramática de valências para o português* de Borba (1996) e *The generative lexicon* de Pustejovsky (1995) e será arrolado o que de cada trabalho se aproveitou para a criação do presente modelo de representação e aquilo em que o modelo se distingue. Para tanto, somente serão expostas as idéias de outras propostas usadas diretamente neste modelo de representação, sendo também apresentados dos tópicos não aproveitados desses trabalhos apenas aqueles que foram deliberadamente excluídos, por apresentarem ou limitações para a busca orientada a idéia, ou por serem inconsistentes com o modelo. Dessa maneira a apresentação de outros trabalhos não se demorará em detalhes que não tenham relação direta com os assuntos desenvolvidos nesta pesquisa. Indicações bibliográficas são acrescidas para o leitor que desejar mais informações sobre os trabalhos comentados.

2.1
A gramática de valências

O trabalho sobre gramática de valências que influenciou este modelo de representação é o encontrado em Borba (1996), cuja obra apresenta características próprias e mais interessantes que outras vertentes do gênero dos estudos da valência gramatical do Português[1].

A mais grata e instigante idéia desse livro é a de que as palavras selecionam a classe e a semântica de seus complementos. Ao esquema de seleções gramaticais e semânticas que uma palavra regente faz de seus argumentos dá-se o nome de valência, termo retirado da Química. Para Borba essa seleção recai na palavra a que corresponderá o papel de núcleo de um sintagma. Esse ponto – a exclusão da

[1] Estamos nos referindo a Vilela (1992).

referência ao sintagma – será a principal diferença entre o tratamento da sintaxe em nosso trabalho e aquele apresentado por Borba (1996).

Uma característica da gramática de valências é que ela centraliza a relação argumental ao aproximá-la da sintaxe. Mas, de forma diferente, Borba não equipara sintaxe à valência como precisamos fazer em nosso modelo.

É importante saber que, em Borba, as valências só se aplicam a palavras que possuem argumento. Para ele, é argumento um sujeito ou um objeto em relação a um verbo, um substantivo em relação a um adjetivo, um complemento nominal em relação a um substantivo. O que fizemos em nosso modelo foi estender a noção de valência para toda e qualquer palavra lexical que estiver exercendo função sintática. Além disso, consideramos que todas as palavras gramaticais sempre participam das valências como elementos de intermediação, o que, de certa forma, ele já fazia com as preposições.

Entendemos que toda função sintática estabelece uma relação de regência, pois toda função sintática precisa que haja outras palavras obedecendo a esquemas determinados para essa função existir. É por isso que estendemos para toda função sintática a noção de argumento e, por isso, podemos aplicar o princípio da valência a qualquer função sintática.

Essa escolha se deveu a dois fatores, o principal deles aproveitar a seleção semântica (ou valência semântica) nas estruturas de conceitos de nosso modelo. O outro fator motivador foi a concisão teórica, já que acabamos usando um princípio único (as valências) para analisar toda a sintaxe.

A concisão teórica também foi a principal razão para não fazermos uso da noção de sintagma, pois queríamos estabelecer o mesmo elo de ligação nas estruturas de conceitos e nas estruturas de argumentos. Por isso, utilizamos o grafo também na estrutura de argumentos, exigindo assim que se trabalhe com um único elemento, a palavra.

Além disso, avaliamos como concisão teórica considerar que a valência sempre selecionará palavras. Ocorre em Borba que ora a valência seleciona palavras (no caso dos adjetivos) ora seleciona sintagmas (no caso dos verbos).

O recurso ao grafo foi também motivado pela pesquisa de Borba. O teórico já vinha estudando o uso de grafos em teorias gramaticais desde seu *Introdução aos estudos lingüísticos* de 1967. Apesar de desenvolvermos um arranjo particular de grafos em nosso modelo, que não guarda nenhuma semelhança com os esquemas de Borba, nossa inspiração inicial provém dessa fonte.

Uma das características mais inovadoras da gramática de valências é a proposta de não existir hierarquia entre os complementos verbais. Sujeito e objetos são vistos ambos como valências do verbo. Mas, ao definir o sintagma como unidade de análise, Borba termina mantendo a hierarquia existente entre uma valência verbal e uma valência nominal, já que o verbo é o centro da sua sintaxe.

Esse fato acaba sendo uma limitação da teoria para nossos fins, já que ela poderia se valer muito mais dos esquemas compartilhados entre as palavras que mudam de classe, como verbos e substantivos deverbais (que para nós devem compartilhar da mesma representação, já que constituem paráfrases[2]).

Borba inova ao considerar que o substantivo é o complemento do adjetivo e não o contrário. Dessa maneira considera que o adjetivo da sentença "casa paterna" está selecionando as propriedades semânticas do substantivo "casa". Mas não considera o sintagma preposicional no esquema das valências. E, dessa forma, na sentença "casa do pai", absolutamente próxima a "casa paterna", o substantivo "casa" não é visto como parte da valência do substantivo "pai". No entanto, para nossos fins, não podemos perder essa relação e, ao contrário, incluímos o sintagma preposicional no esquema de valências. Por isso a gramática de Borba é bem menos "valencial" que nosso modelo de representação.

Outra característica que influenciou nosso modelo foi a utilização do esquema valencial para a determinação do sentido de uma palavra, ainda que a valência semântica seja tratada diferentemente em ambos os modelos. Em Borba, são selecionados traços semânticos que os complementos deverão possuir. Já em nosso trabalho, o que se seleciona é um subgrafo da estrutura de conceito do complemento.

[2] Essa equivalência que fazemos se assemelha muito com o compartilhamento de estrutura profunda entre verbos e substantivos deverbais em Chomsky (1965: 16-18).

Ainda que importante em seu trabalho, Borba não sistematiza muito detalhadamente a semântica no processo de seleção.

Ainda existem outros pormenores que compartilhamos com a teoria de Borba. A maioria dos nomes de classes gramaticais foi retirado de Borba: por exemplo, "qualificador" e "substantivo-evento". No entanto, essas classificações servem para denominar conjuntos diferentes de palavras além de possuírem definições um tanto distintas[3].

Ao extremar a gramática de valências para atender nossos propósitos, tivemos que acrescentar as palavras de ligação como fazendo parte da valência da palavra regente. Borba faz isso para as preposições. O modelo usado para a busca orientada a idéia pretende considerar como elemento de ligação uma série de outras palavras e, por causa disso, acaba considerando como palavra de ligação (e mesmo como palavra gramatical, o que é ainda mais extremo) artigos, pronomes, verbos de ligação, verbos modais e auxiliares.

Por exemplo, para duas sentenças como "o violão quebrado" e "o violão está quebrado" serem vistos como paráfrases (ou, pelo menos, para levarmos em conta que compartilhem boa parte de suas idéias), devemos considerar que, na segunda sentença, o adjetivo "quebrado" deve continuar selecionando o substantivo "violão" como seu argumento. Nesse caso, o verbo de ligação funcionaria como elemento de ligação. E mais, se esse verbo funciona como um elemento de ligação, ele deverá ser visto pelo modelo como palavra gramatical. O mesmo ocorre com verbos modais e auxiliares (como os verbos das sentenças "o menino *deveria* sair de casa" e "o menino *vai* sair de casa"), que devem ser considerados como elementos de ligação entre o verbo principal (no caso o verbo "sair") e seus argumentos. Nesse exemplo, bastaria considerar o verbo modal "deveria" e o verbo auxiliar "vai" como elementos

[3] As definições de Borba para as classes de palavras levam em conta, por exemplo, o critério distribucional (que para ele é o lugar que uma palavra ocupa seqüencialmente na sentença). Já no modelo apresentado aqui, o lugar que uma classe de palavra ocupa é a de complemento desta ou daquela palavra. Por exemplo, o nome é a classe de palavra que serve de argumento para o evento e para o qualificador.

de ligação do argumento “menino”. Para o outro argumento, “casa”, bastaria considerar como elemento de ligação a preposição “de”.

Concluindo esse tópico, a idéia de valência é o que fundamenta a estrutura de argumentos de nosso modelo. A diferença básica em relação ao modelo de Borba é que este restringe a atuação das valências a algumas classes de palavras. Nosso modelo de representação lingüística, portanto, é uma gramática de valências extremada, em que toda a sintaxe é explicada pela valência.

2. 2
O léxico gerativo

Foi de Pustejovsky (1995) que retiramos o termo “estrutura de argumentos”. Foi-nos de imensa valia o tratamento semântico em “níveis de representação”, como ele propõe, principalmente por essas representações constarem em uma espécie de dicionário mental, com níveis de representação para cada entrada lexical. Isso significa que cada aspecto de uma palavra (significado, sintaxe e até certos aspectos pragmáticos) é tratado no léxico, e a cada aspecto corresponde um tratamento, o que ele chama de “estrutura”.

Pustejovsky descreve uma entrada lexical como um complexo estrutural: a estrutura de argumentos, a de eventos, a qualia e a de herança. Somente utilizamos uma de suas estruturas – a estrutura de argumentos –, mas criamos outras estruturas aos moldes dos níveis de representação dele, voltadas para nossos propósitos.

Como o próprio título do livro sugere, o modelo de Pustejovsky é gerativo. E isso significa que esse trabalho tenta se integrar ao modelo gramatical do programa gerativo vigente no período. Dessa maneira, o léxico gerativo tenta abarcar uma série de elementos gerativos como o critério-teta e o princípio de projeções em suas estruturas. Mesmo assim, o teórico traz propostas bem diferentes aos do gerativismo do período.

O modelo de Pustejovsky não visa o tratamento semântico para um uso específico como um buscador. Ele visa uma descrição global da língua. Dessa maneira, o modelo traz muitos detalhes não só irrelevantes para o buscador orientado

a idéia como também acaba por resultar em problemas de ordem teórica e de ordem prática para nossa finalidade, como veremos mais a frente.

Apesar de usarmos a terminologia "estrutura de argumentos" de Pustejovsky, nossas conceituações e usos são bem distintos. Para nós, "estrutura" é apenas uma coisa: grafo. Para Pustejovsky, uma estrutura pode ser configurada de maneira diversa, dependendo de qual for o tipo de estrutura e até mesmo dependendo de qual for a entrada lexical.

Não precisamos aproveitar todos os diversos critérios e estruturas de Pustejovsky. Como não temos a mesma finalidade que ele, podemos tratar da mesma forma muitos dos fenômenos tratados por Pustejovsky em diferentes níveis estruturais. A teoria de Pustejovsky é muito complexificada, já que pretende dar conta dos mais diferentes fenômenos lingüísticos. Por exemplo, sua estrutura argumental é subdividida em uma série de propriedades. A entrada seleciona as propriedades semânticas de um argumento (como faz a gramática de valências), mas o argumento também deve possuir uma propriedade definida como "formal". Usando um exemplo de Pustejovsky, a entrada "fazer um bolo" pede um argumento que tenha a propriedade de ser animado e tenha a propriedade formal de ser um objeto físico.

Essa subdivisão gera problemas teóricos e práticos para nossa finalidade. Primeiramente, Pustejovsky trata qualquer palavra com os mesmos critérios. Isso é importante do ponto de vista da concisão teórica, mas as escolhas que faz dos critérios não são concisas o suficiente para nossa proposta. A existência de uma propriedade formal mostra isso. Apenas substantivos que nomeiam coisas concretas com existência material (como as palavras da nossa classe "nome") poderão ter uma propriedade formal. O mesmo acontece com a estrutura de eventos, que só é aplicada a verbos e palavras derivadas de verbos.

Esse problema teórico gera um problema prático, porque, como Pustejovsky atribui uma determinada propriedade a qualquer entrada lexical e, por isso, muitas vezes acaba mudando sutilmente os critérios de uma classificação para poder abarcar palavras que não possuem tal propriedade. Por exemplo, ele atribui a propriedade argumento, da estrutura de argumentos, a palavras que não possuem argumentos. Mas, para isso, ele altera a representação dessa propriedade. Para uma palavra que

possua argumentos, a propriedade deve ser representada pelo atributo semântico que seu argumento deve possuir. Por exemplo, para o verbo “gravar”, Pustejovsky atribui dois argumentos, um com a propriedade “objeto físico” e o outro, com “informação” (1995: 129). Já para uma palavra sem argumentos, essa propriedade deve ser representada pelo próprio atributo semântico da entrada lexical. Por exemplo, Pustejovsky representa a propriedade argumento de “faca” como “ferramenta”.

Temos mais um grande problema nessa classificação para podermos aplicá-la a nosso modelo. Pustejovsky define o atributo semântico de palavras que não possuem argumento, como é o caso de “faca”. Realmente, como argumento de um verbo como “cortar”, “faca” seria apropriadamente definida como uma “ferramenta” . No entanto, em nosso modelo bastaria atribuirmos a “faca” a propriedade “objeto físico”. Isso porque, para “faca” constar como argumento de um verbo como “cair”, não é necessário saber que esta palavra possui o atributo semântico “ferramenta”, mas seria suficiente saber que possui o de “objeto físico”.

Temos aí outro problema no modelo de Pustejovsky que dificulta sua aplicação para nossos propósitos: a falta de integração entre as propriedades. Ora, no caso da propriedade formal vista acima, não existe relação entre um objeto físico e um ser animado? Todo ser animado não é um objeto físico? Portanto, quando uma entrada seleciona a propriedade “ser animado”, ela já estaria selecionando a propriedade “objeto físico”. Dessa maneira, para o nosso modelo, a propriedade formal é redundante com a propriedade semântica além de ser desnecessária por não ser aplicada a qualquer tipo de palavra.

Outro exemplo que tornaria nosso modelo redundante se aplicássemos diretamente o modelo do léxico gerativo pode ser visto na propriedade télica da estrutura qualia. Pustejovsky atribui à palavra “faca” a propriedade télica “cortar” (1995: p.129). Na realidade, ele atribui uma cláusula de Horn a essa propriedade, para poder atribuir todos os papéis argumentais de “cortar” e em quais deles entraria “faca”. A propriedade télica de “faca” é redundante com a estrutura argumental de “cortar”, uma vez que bastaria a representação dessa relação em uma das entradas lexicais.

Optamos por tratamento bem diversificado ao de Pustejovsky. Escolhemos um único critério para definir nossas estruturas e também escolhemos estruturas que valessem para todas as palavras. É claro que existe um grupo de palavras para as quais não valem essas estruturas: as palavras gramaticais. No entanto, essas palavras não são tratadas pelas estruturas, enquanto todas as palavras lexicais são tratadas pelas estruturas.

As idéias apresentadas por Pustejovsky, apesar de não atenderem nossa finalidade por completo, são válidas para satisfação de sua proposta: a descrição da língua. E, nesse intento, são muito instigantes e criativas. A idéia de tratar as entradas lexicais a partir de vários níveis de representação mudou por completo os rumos de nossa pesquisa. Por isso, foi feita referência ao estudo de Pustejovsky, ao nomearmos nossas estruturas como ele faz.

Pustejovsky tem uma percepção muito aguçada de fenômenos lingüísticos. Acredito que suas descobertas sobre as estruturas de eventos possam ser usadas para resolver nosso problema latente com os verbos que pressupõem outros eventos como é o caso do verbo "empatar". Ele percebeu que alguns eventos pressupõem outros e, mais interessante, pressupõem uma ordem temporal entre esses eventos. No nosso caso, "empatar" pressupõe dois eventos "marcar gol", sendo que um é anterior ao outro, além de terem como argumentos times diferentes. Ainda falta no nosso modelo integrar essas estruturas de eventos às outras estruturas.

2. 3
Representações de conceitos

Nesse tópico, vamos apresentar três formas usadas ora para representar a semântica (traços semânticos) e ora para representar o conhecimento (*frames* e redes semânticas). Essas formas de representação foram criadas para usos diferentes do nosso: os traços semânticos nasceram dos estudos lingüísticos e visavam integrar semântica à sintaxe; os *frames* e as redes semânticas foram criados pela inteligência artificial, visando maneiras de atrelar a um programa um conhecimento específico que pudesse manipular. O que essas formas de representação têm em comum é que

são formas de representar conceitos e relações entre conceitos. E foi isso que nos interessou nessas representações: seriam elas boas formas para representar nossos conceitos? Veremos que somente foi aproveitada dessas representações a idéia básica de usar conceitos e relações entre conceitos, mas o relacionamento entre os conceitos e até mesmo o que consideramos como sendo um conceito e uma relação é algo bem diferente.

2. 3 .1
Traços semânticos

O modelo dos traços semânticos (Pietroforte&Lopes, 2003: 118-119) é uma aplicação do esquema de traços da fonética à representação dos significados das palavras. Na fonética, tínhamos um conjunto de pouco mais de meia dúzia de pontos de articulação possíveis de serem executados. Um fonema se distinguiria de outro por apresentar uma combinação única de pontos de articulação que co-ocorreriam. Para cada ponto de articulação executado pelo fonema, dizia-se que este fonema teria traço positivo para aquele ponto de articulação, e para os não executados, o fonema teria traço negativo.

Para a semântica, aproveitou-se essa idéia, e as palavras teriam traços positivos e negativos para conceitos. Basicamente esses conceitos representariam parte do significado de uma palavra. Os traços seriam atribuídos para contrapor palavras que se relacionariam semanticamente, como os antônimos, ou que pudessem servir como argumentos de outras palavras.

Por exemplo, uma palavra como “garota” teria os traços [+humano, +feminino, +jovem], enquanto homem teria os traços [+humano, –feminino, –jovem]. Esse sistema binário pouparia conceitos, por exemplo, ao excluir um conceito “masculino”, sendo entendida a idéia de masculino pelo traço negativo para o conceito “feminino”.

No caso de traços atribuídos na seleção de argumentos, temos que uma palavra selecionaria de seu argumento um traço positivo ou negativo determinado. Somente poderia configurar como argumento uma palavra que contivesse o mesmo traço

pedido. Por exemplo, o verbo "vender" pediria o traço [+humano] para o seu sujeito, enquanto o verbo "pular" pediria o traço [+animado]. Então uma palavra como "José", que possui os traços [+humano, +animado], poderia se configurar como sujeito de ambos os verbos, enquanto "gato", cujos traços são [–humano, +animado], somente poderia se configurar como sujeito de "pular", e uma palavra como "pedra", com os traços [–humano, –animado], não poderia se configurar como sujeito de nenhum dos verbos.

São vários os problemas que nos fizeram não optar por esse modelo de representação de conceitos. Uma das maneiras como esse modelo escolhe os conceitos das palavras é contrapondo palavras do mesmo campo semântico e escolhendo traços que representariam pequenas diferenças de significado entre elas. Por exemplo, entre "cadeira" e "sofá", a primeira teria os traços [–vários assentos, –estofado] enquanto a segunda teria [+vários assentos, +estofado]. Para a finalidade de busca orientada a idéia, essa estratégia de atribuição de conceitos cria conceitos desnecessários e, às vezes, cria conflitos entre conceitos. Isso porque o computador não precisa ser informado de que uma cadeira pode ter braços, que tem somente um lugar, que não é estofada etc., se somente interessa saber que ela não pode ser reescrita como "sofá". Isso explica o caráter desnecessário dessa estratégia.

O problema do conflito entre conceitos acontece porque a orientação a idéia não trata apenas de palavras que compõem um mesmo campo semântico. É possível que expressões contenham idéias que podem ser atribuídas a outro conjunto de palavras, ou mesmo a uma única palavra. Dessa maneira, os conceitos das várias palavras de uma sentença precisam ser unidos. Assim, uma frase como "a cadeira estofada" teria um conjunto conflitante de traço [–estofado], vindo de "cadeira", e [+estofado] vindo de "estofada".

Preferimos determinar conceitos entre as palavras de um mesmo campo semântico pela forma como elas formam paráfrases e não pelo entendimento que fazemos de seus significados.

Já com relação à estratégia de atribuir conceitos através da seleção semântica dos argumentos, esbarramos em outros problemas. Vejamos o exemplo apresentado

anteriormente, em que o verbo "vender" pede o traço [+humano] e o verbo "pular" o traço [+animado].

Temos pelo menos dois problemas aí. O primeiro é que os conceitos não se relacionam entre si. Por exemplo, podemos supor que toda palavra que contiver o traço [+humano] será [+animado]. O segundo ponto é que, para impedir que uma palavra como "pedra" seja aceita como argumento desses verbos, é necessário que se atribua a ela os traços [–humano, –animado]. Assim, mesmo que uma palavra não possuísse um dado conceito, deveria ter o traço negativo para aquele conceito. Estendendo isso para toda a língua, é presumível imaginar que a quantidade de traços que participam de todas as seleções argumentais possíveis não é pequena, e que cada entrada do léxico teria de conter todo esse enorme inventário de traços semânticos. Um sistema computacional que fizesse tal verificação de traços seria inviável, ou pelo menos bem pouco prático.

Por isso, excluímos a idéia de traço positivo ou negativo para um conceito. Ou uma palavra possui um conceito ou não possui. E, para que todos os demais problemas sejam resolvidos, os conceitos têm que estar relacionados entre si. Os outros modelos apresentam formas de relacionar os conceitos.

2. 3. 3
Redes semânticas e *frames*

A rede semântica (Araribóia, 1988: 228) organiza os conceitos com que trabalha através de arcos que unem os conceitos que estabelecem entre si alguma relação. Essa relação pode ser de várias naturezas, sendo assim, cada seta que representa uma relação será especificada com o tipo de relação que ligam os conceitos. Eis um exemplo de rede semântica:

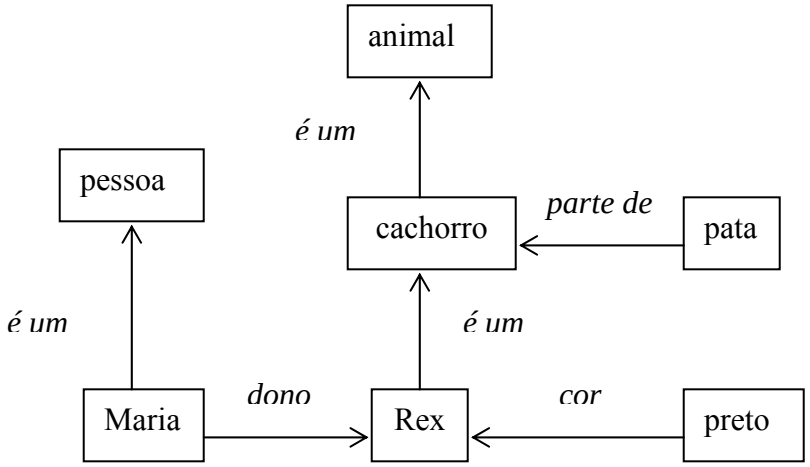

Figura 1 – rede semântica

Como se vê, esse esquema poderia ser representado por cláusulas de Horn de tal forma: "é_um(Maria, pessoa)", "é_um(Rex, cachorro)", "é_um(cachorro, animal)", "dono(Maria, Rex)", "parte_de(pata, cachorro)", "cor(preto, Rex)".

A objeção que fazemos à rede semântica (e também às cláusulas de Horn) para nossa proposta é que ela não é a mais adequada para se trabalhar com a semântica de palavras. O mesmo argumento que usaremos para explicar essa objeção é também válido para contestar o uso dos *frames*, por isso apresentarei os argumentos para essa objeção mais adiante.

Para nossos propósitos, basta entender que os *frames* (Araribóia, 1988: 231 − para maiores detalhes) são derivados das redes semânticas, mas têm uma configuração um pouco diferente. Todo conceito é armazenado num escaninho. Em cada escaninho são atribuídos ao conceito atributos e eventos. Entre os escaninhos somente são atribuídas as relações *é um*. Dessa maneira, um conceito derivado de outro pela relação *é um* herda as propriedades e eventos do conceito "pai". O conceito de programação orientada a objetos é bastante semelhante ao modelo de *frames*.

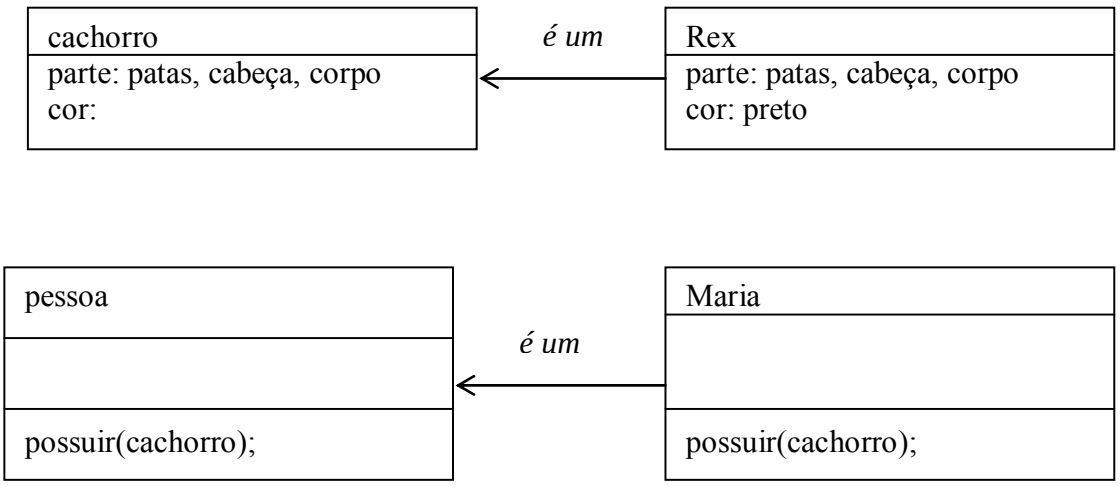

Figura 2 – modelo de *frames*

O problema de *frames* e redes semânticas, para uma aplicação lingüística, é que existem idéias que são conceitos e idéias que são relações entre conceitos. A idéia *é um* é vista como uma relação enquanto "pessoa" é um conceito.

Na verdade tanto *frames* quanto redes semânticas são aplicados a conhecimentos de mundo, o que justifica essas distribuições. O nosso modelo não é uma representação do conhecimento, aliás, uma representação da semântica de textos terá dificuldades ao ser tratada através de modelos de representação do conhecimento como os *frames* e redes semânticas.

Para nos referirmos às idéias e às abstrações que formamos a partir de nosso entendimento do mundo, usamos palavras. Usamos palavras para nos referirmos a entidades com existência física, como cachorro e Maria. Mas também usamos palavras para nos referirmos a relações abstratas como posse, parte. Essas relações também podem ser atribuídas através do relacionamento sintático entre as palavras: "Rex é um cachorro", "o cachorro Rex". Nesses dois exemplos, através das relações sintáticas pode-se estabelecer entre as palavras a relação *é um* usada pelo *frame* e pela rede semântica.

Idéias como *posse* podem estar distribuídas em palavras que sejam substantivos ("dono") ou verbos ("possuir"), ou ainda numa relação sintática ("cachorro de

Maria"). Por isso, não achamos que seja possível tratar uma palavra ora como uma relação e ora como um conceito. E, além disso, as palavras podem conter mais de um conceito e existirem ainda relações entre esses conceitos.

Por isso, em nosso modelo somente existe uma relação que os conceitos podem estabelecer entre si: a relação de especificação[4]. Todas as outras relações das redes semânticas e dos *frames* são tratados como conceitos. Dessa forma, o que nas redes semânticas e nos *frames* é um conceito, em nosso modelo é um conceito. O que numa rede semântica é relação e num *frame* é atributo ou evento ou herança, em nosso modelo também é um conceito.

Concluindo, nosso conhecimento do mundo obviamente influencia e, talvez, gerencie os significados das palavras e textos. Mas isso não significa que para representarmos a semântica precisamos sistematizar o conhecimento do mundo. Para o uso da busca por idéias, a representação do conhecimento é dispensável.

2. 4
A *web* semântica

A *web* semântica (Moura, 2004) não é exatamente uma influência no nosso trabalho, mas como ela é das poucas tecnologias em desenvolvimento que se propõe a resolver nosso problema (a busca por idéias e não por palavras) é indispensável mencioná-la.

Foram encontradas duas grandes limitações na *web* semântica que, por serem o cerne dela, nos fizeram optar por pensar numa solução completamente diferente para a busca orientada a idéia, ao invés de adaptar essa tecnologia para a nossa finalidade. Os problemas são a) o esquema de ontologias, e b) o recurso a *tags* para indexar os conceitos.

A ontologia é uma combinação de redes semânticas e *frames*. Ela é descrita por conjuntos distintos de conceitos e relações, um conjunto desses conceitos que

[4] Note que especificação é uma meta-conceituação, não é a palavra "especificação" nem os conceitos que possam definir a idéia que "especificação" possui. Como meta-conceituação, especificação somente guarda ligação com "especificação" numa realidade externa ao modelo. A relação de especificação não cai no mesmo problema que *é um* ou *parte de* caíram, porque a palavra "especificação" terá seus conceitos, e palavras ou estruturas gramaticais que contiverem a idéia que "especificação" possuir, compartilharão dos conceitos que "especificação" possui.

estabelecem relações hierárquicas, um conjunto de conceitos que estabelecem outros tipos de relações, e regras para o tratamento desses conjuntos.

Nosso argumento contra essas subdivisões para o tratamento da semântica é válido para a *web* semântica, uma vez que os documentos encontrados em sua grande maioria na internet são textos. Textos são compostos de palavras e relacionamento sintático entre palavras. É complicado tratar palavras e sintaxe ora como conceitos, ora como relações.

São utilizadas de forma predominante na *web* semântica as *tag*s do XML (eXtended Markup Language). Uma *tag* no XML é qualquer informação especificada entre os sinais de menor ("<") e maior (">") atribuídas a uma extensão de texto compreendida entre o lugar onde essa *tag* é inserida até onde é inserida outra *tag*, sinalizando o fim da atuação dessa informação. Veja o exemplo:

Frase original:
Maria tem um cachorro.

Frase indexada:
<posse> <pessoa> Maria </pessoa> tem um <animal>cachorro </animal></posse>

Qual o problema desse tipo de indexação? A *tag* atua numa extensão entre uma *tag* que indica o começo, até outra que indica o fim, isso é, um conceito é válido para uma seqüência de palavras. O problema é que as palavras não se relacionam somente em seqüência. Muitas das idéias são formadas não apenas por uma palavra, mas pelo relacionamento de algumas palavras, e, por isso, um significado não pode ser atribuído somente de maneira seqüencial.

Apesar do modelo de *tag*s permitir que uma *tag* possa vir intercalando outra, ele não permite que parte da seqüência de palavras entre *tag*s não seja especificada por essas *tag*s. Por exemplo, se numa sentença como "Maria comprou um lindo vestido" indexarmos como "<comprar> Maria comprou um lindo vestido </comprar>", estaremos admitindo que "lindo" faça parte da idéia de "comprar".

Isso pode parecer pouco relevante, mas não é. Imagine que uma idéia somente seja expressa no relacionamento entre duas palavras. Imagine agora que essas palavras apareçam bem distantes no texto, e que, através de anáforas, possamos atribuir a relação entre elas e, portanto, o significado mencionado possa ser

estabelecido. Pelo modelo de *tag*s a idéia será atribuída a todo o trecho entre as palavras. E isso pode se extender a várias linhas.

Entendemos que a *web* semântica quis aproveitar um modelo pré-existente – o XML –, adaptando-o ao problema que pretende solucionar. Mas achamos mais viável modelar uma solução que atenda a própria configuração dos objetos tratados. Dessa maneira, a estrutura dos conceitos se assemelha à estrutura sintática, pois é através da sintaxe que as palavras se relacionam.

2. 5
A Universal Networking Language (UNL) e a dependência conceitual

Esses foram dois modelos que influenciaram as bases iniciais de nossa pesquisa, embora, de certa forma, seus princípios principais tenham sido abandonados durante o amadurecimento de nosso modelo de representação lingüística.

O modelo da UNL –Universal Networking Language – (Specia&Rino, 2002) nos influenciou e continua influenciando pela idéia de interlíngua adaptada a representação lingüística. A interlíngua vista como representação pela UNL não é uma língua franca, como pode ser entendida em seu sentido mais lato. Na UNL, a interlíngua age como uma língua intermediária entre duas traduções, isto é, se dois textos ou duas sentenças são traduções, então eles devem ter a mesma representação na interlíngua. A idéia por trás disso é que todas as línguas poderiam ser representadas pela mesma interlíngua e, assim, a tradução seria muito facilitada.

Utilizamos também o princípio da língua intermediária, que em nosso modelo é a estrutura de conceitos. Mas, ao invés de essa interlíngua representar sentenças que são traduções, representam sentenças que são paráfrases. Em suma, podemos dizer que a estrutura de conceitos é uma língua intermediária entre dois textos que são paráfrases.

No entanto, termina aí a influência. Isso porque, no momento em que a UNL pretende ser uma única representação, servindo a toda e qualquer língua, ela tem que admitir uma concepção teórica que considera que existem relações universais entre o conteúdo semântico de todas as línguas. E é o que ela propõe. A UNL se baseia em conceitos primitivos compartilhados por todas as línguas.

Isso por si só não seria um problema, mas para nosso uso específico, levantar esses universais é um trabalho dispendioso e desnecessário. É mais fácil e rápido, em tempo de implementação humana, atribuir os conceitos mecanicamente às paráfrases, e é mais eficiente e exige menos capacidade de processamento, em tempo de execução do computador, limitar-se somente às paráfrases.

Formalmente a UNL sofre com o que tem sido o problema de quase todos os modelos analisados aqui para nosso uso: proliferam análises onde a concisão poderia predominar. Acreditamos que isso acontece quando os modelos teóricos pretendem explicar todos os usos, todas as vicissitudes, todas as manifestações da língua, e não se limitam a buscar unicamente o tratamento daquela parte do sistema lingüístico suficiente para solucionar o problema proposto ou as questões levantadas.

A busca por uma verdade total válida para toda a integridade da língua logra, muitas vezes, a teoria.

O modelo de Roger Schank (Schank&Tesler, recurso eletrônico), a dependência conceitual, também se baseia no princípio das primitivas. Para ele, o significado de um evento seria composto por uma combinação própria de conceitos primitivos, sendo que o número total de conceitos primitivos e disponíveis para compor um significado não passaria de mais ou menos uma dúzia.

Essa idéia a princípio motivou o início de nosso trabalho, inclusive sendo apresentada em nosso projeto de pesquisa. Mas tivemos de desistir da idéia de atribuir conceitos primitivos à estrutura de conceitos das palavras e sentenças.

A primeira razão para essa desistência foi que quisemos aplicar as primitivas não apenas aos eventos, mas a todas as palavras lexicais, já que a idéia era utilizar a decomposição de conceitos para as paráfrases. O problema das primitivas para nosso uso é que ela se baseia no entendimento de uma idéia. Por exemplo, existe uma primitiva para movimento e outra para agente, porque várias palavras têm a idéia de movimento e agente. As palavras que contiverem essas idéias deveriam ter essas primitivas em suas composições.

No entanto, compor o entendimento que fazemos de um substantivo concreto, seja através de conceitos primitivos ou não, é impossível. Por exemplo, ao levantar as idéias que compõe uma banana , podemos dizer que ela é uma fruta, que tem a casca

amarela com manchas pretas, que tem a casca verde quando não está madura, que tem a casca preta quando está passada, que ela é branca (ou levemente amarelada) em sua polpa, que a polpa é macia e doce, que tem pequenas sementes pretas, que tem forma ablonga levemente curvada, que é cilíndrica, que a casca é grossa, que podemos descascá-la com as mãos...

Como se vê, podemos prosseguir infinitamente na enumeração dessas propriedades. E nunca elas serão suficientes. Se uma pessoa nunca tiver visto nem comido uma banana na vida, provavelmente a imagem mental que ela faria com essa enumeração estaria longe de ser uma banana real. E mais, para cada uma das pequenas explicações que demos à banana, temos que notar que deveríamos transcrevê-las em conceitos primitivos. Então também teríamos que definir em primitivas o que é fruta, o que é casca, o que é amarelo, o que é macio, etc. Assim, o tamanho da composição de um significado de uma fruta, mesmo que em primitivas, poderia ser algo imenso e inviável.

A teoria da dependência conceitual possui outras características que nos influenciaram. Foi com ela que percebemos que os conceitos deveriam se relacionar. Mas no modelo de Schank os relacionamentos são bem mais complexos que os nossos.

Sua formalização das relações entre conceitos, como a nossa, se apresenta de forma gráfica, em que as relações são representadas por setas. Porém, à moda das redes semânticas, no modelo de Schank existem várias relações entre os conceitos, e para cada tipo de relação corresponde um desenho de seta diferente.

A complexidade de desenhos é empolgante no princípio, por parecer que dá conta da totalidade de problemas que procura resolver. Mas por que não devemos confundir uma teoria complexa com uma teoria completa, tivemos que escolher trabalhar com um único tipo de relação e apenas uma representação para ela. O modelo da dependência conceitual sofre dos mesmos problemas relativos aos vários tipos de relações nas redes semânticas. E, além disso, as partes do sistema de Schank não interagem bem, principalmente pela falta de um estudo da relação gramatical entre as palavras, porque, afinal, estudar as palavras não era sua intenção exatamente.

2. 6
Conclusão do capítulo

A primeira conclusão que podemos tirar ao estudar os modelos não lingüísticos é que eles não tratam de palavras, tratam das idéias que "existem" no mundo, mesmo se servindo das palavras muitas vezes como objeto de pesquisa. Esses modelos confundem muitas vezes as palavras com as idéias que elas representam, como uma espécie de língua de Adão.

Mas essa confusão não é um problema para os teóricos desses modelos, afinal eles não estudam a língua. Para seus fins a confusão não é má.

A confusão se torna problemática nos estudos que se utilizam desses modelos para tratar da língua. Ao confundir idéia e palavra, esses modelos, ao serem aplicados a problemas lingüísticos, deixam sem solução muitas questões e trazem outras tantas questões desnecessárias por fugirem do âmbito lingüístico.

Com relação aos demais trabalhos apresentados, nosso modelo visa tratar um problema ainda não resolvido. Dessa maneira as técnicas e teorias existentes para tratar de outros problemas só foram usadas como inspiração para a solução de nosso problema.

Ao invés de pesquisar, testar e adaptar técnicas e teorias existentes para resolver um problema inédito, resolvemos partir do zero e criar uma solução também inédita, mesmo sabendo que inspirações e *insights* advindas de outros conhecimentos, além de serem inevitáveis, são muito bem-vindas.