Pontifícia Universidade Católica
do Rio de Janeiro

Carolina Monte Ferreira Gonçalves

As bases lingüísticas para a busca orientada a idéia

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós-graduação em Letras da PUC-Rio como requisito parcial para a obtenção do título de Mestre em Letras.

Orientadora: Violeta de San Tiago Dantas Barbosa Quental

Rio de Janeiro, março de 2006

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Carolina Monte Ferreira Gonçalves

As bases lingüísticas para a busca orientada a idéia

Dissertação apresentada como requisito parcial para a obtenção de grau de Mestre pelo Programa de Pós-graduação em Letras do Departamento de Lestras do Centro de Teologia e Ciências Humanas da PUC-Rio. Aprovada pela Comissão examinadora abaixo assinada.

_____
Profa. Violeta de San Tiago Dantas Barbosa Quental
Orientadora
Departamento de Letras – PUC-Rio

_____
Profa. Maria Carmelita Padua Dias
Departamento de Letras – PUC-Rio

_____
Profa. Rove Luiza de Oliveira Chishman
UNISINOS

_____
Prof. Paulo Fernando Carneiro de Andrade
Coordenador Setorial do Centro de Teologia e
Ciências Humanas – PUC-Rio

Rio de Janeiro, ____ de _____de _____.

Carolina Monte Ferreira Gonçalves

Graduou-se em Letras Português Literaturas na Universidade Estadual do Rio de Janeiro em 2003.

Ficha Catalográfica

Gonçalves, Carolina Monte Ferreira

As bases lingüísticas para a busca orientada a idéia / Carolina Monte Ferreira Gonçalves ; orientador: Violeta de San Tiago Dantas Barbosa Quental. – Rio de Janeiro : PUC, Departamento de Letras, 2006.

132 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras.

Inclui referências bibliográficas.

1. Letras – Teses. 2. Lingüística computacional. 3. Recuperação de Informação. 4. Teoria sintática. 5. Teoria Semântica. I. Quental, Violeta de San Tiago Dantas Barbosa. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

Para Leda

Agradecimentos

Ao velho Omolu. *Valei-me, meu pai, atotô, Obaluaê*.

À minha mãe, Kátia, ao Pedro e ao Rafa, e aos demais familiares – em especial ao Dani pela tradução do resumo.

Ao meu vô Monte (em memória), cuja pequena biblioteca me iniciou nos estudos lingüísticos.

À Ledinha pelo companheirismo, carinho e paciência.

Aos Justas Juliana, Marcelo e ... Edmar pelos projetos megalomaníacos.

Aos amigos Fernando e Carlos pelas conversas e diversões.

À minha orientadora Violeta Quental pelas importantes contribuições.

Aos colegas e professores da Puc-Rio.

Aos funcionários do Departamento de Letras, em especial à sempre prestativa Chiquinha.

Ao CNPq e à Puc-Rio pelos auxílios concedidos.

Resumo

Gonçalves, Carolina Monte Ferreira. **As bases lingüísticas para a busca orientada a idéia**. Rio de Janeiro, 2006. 132p. Dissertação de Mestrado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

A busca orientada a idéia é um novo paradigma para mecanismos de busca em acervos compostos por arquivos de texto. Esse paradigma visa resolver um problema comum aos mecanismos de busca: exigir que o usuário preveja as palavras contidas nos documentos que possam conter a informação que procura, impossibilitando-o, assim, de se concentrar diretamente na informação desejada. Buscando solucionar esse problema, são propostas as bases lingüísticas para o desenvolvimento de um modelo teórico preliminar que acrescente dados semânticos aos arquivos de texto. Nesse modelo, a informação semântica de um texto é representada através do que se chamou de estrutura de conceitos. O principal intuito das estruturas de conceitos é representar de uma mesma maneira frases que expressem o mesmo significado, ou seja, as paráfrases apresentam a mesma estrutura de conceitos. Serão expostos nesta dissertação os primeiros elementos do modelo em suas partes semântica, sintática e textual, além da integração entre as mesmas. A dissertação apresenta ainda um estudo de caso a fim de exemplificar o desenvolvimento de uma aplicação para busca de arquivos de texto em que essa tecnologia seria usada.

Palavras-chaves

Letras; lingüística computacional; recuperação de informação; teoria sintática; teoria semântica.

Abstract

Gonçalves, Carolina Monte Ferreira. **Linguistic basis for idea-oriented search**. Rio de Janeiro, 2006. 132p. Dissertação de Mestrado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

The idea-oriented search is a new pattern for search engines whose databases are composed by text files. This pattern sets out to solve a usual problem for search engines: demanding that users foresee which words are contained in the desired document, keeping them from focusing on the information they are indeed seeking. To solve this problem, the linguistic bases for the development of a theoretical model that can add semantical data to the text files are laid down. In this model, the semantical information of a text is represented by what has been referred to as structure of concepts. The main goal of the structure of concepts is to give one single representation to sentences that express the same meaning. Thus, paraphrases present the same structure of concepts. In this dissertation, the first elements of this model are exposed in its semantical, syntactic and textual parts. Also present are the integration of these elements. A small case study is presented as well, with the intention of illustrating the development of an application for text files databases search engines in which this technology is used.

Keywords

Computational Linguistics; information retrieval; syntactic theory; semantic theory.

Sumário

Lista de Figuras

1
Motivação

A presente pesquisa apresenta as primeiras hipóteses para o desenvolvimento de um novo padrão para mecanismo de buscas de arquivos de texto. Através desse mecanismo, o usuário escreverá a informação que ele quer encontrar e o programa lhe retornará os documentos em que essa informação apareça, sem que seja necessário ao usuário prever quais palavras o autor do texto usou para expressar a informação requerida. A esse novo padrão de busca dei o nome de orientado a idéia, em contraposição ao padrão mais comum de busca que estamos chamando orientado a palavra.

A área de conhecimento em que essa pesquisa se insere é a da Recuperação de Informação (RI), para a qual advogamos a necessidade de conhecimento lingüístico. A Recuperação de Informação normalmente trata de duas problemáticas: i) a identificação de um documento dentre outros num acervo, a partir do pedido de um usuário na forma de texto, e ii) a catalogação dos documentos no acervo conforme assunto, resumo ou outros critérios de indexação.

Para nós só será interessante tratar a primeira problemática.

Normalmente a recuperação de informação se limita a recuperar arquivos e não informações (Ferneda, 2003). Mas muitas vezes é informação aquilo que o usuário pensa estar recuperando, ou que gostaria de estar recuperando. Portanto, recuperar a informação é o requisito primordial a ser atendido pelo mecanismo de busca.

Essa limitação gera a imensa dificuldade que o usuário tem em tentar adivinhar quais palavras deve usar para conseguir o resultado desejado. Imagine-se uma situação corriqueira como usar o terminal de computador de uma biblioteca que rode um mecanismo de busca para os livros de seu acervo. Se o usuário resolver encontrar um livro que contenha uma determinada informação, terá muita dificuldade para descobrir com que palavras essa informação terá sido classificada pelos bibliotecários, se é que a informação foi catalogada.

Essa dificuldade fica bem mais evidente quando os arquivos buscados não se limitam a itens de um banco de dados, como no exemplo da biblioteca, mas quando o documento procurado é um arquivo de texto. As páginas HTML da internet são

exemplos disso. Os arquivos de texto permitem um campo mais profícuo para a Recuperação de Informação, pois a busca pode recair não apenas nos recursos usados para a catalogação (títulos, resumos ou outras formas de indexação), mas, como o próprio conteúdo do arquivo é composto por palavras, tal conteúdo também pode ser alvo do mecanismo de busca.

Nossa pesquisa se limita a procurar soluções, com base em informações lingüísticas, para buscadores cuja pesquisa se faz sobre documentos do tipo arquivo de texto.

1.1
A busca orientada a palavra

Os buscadores de páginas da internet não recuperam informação, recuperam palavras. O mecanismo de busca "enxerga" uma página simplesmente como uma seqüência de palavras (entendidas como seqüências de caracteres) e sua tarefa é encontrar os documentos que contenham as mesmas palavras que o usuário requisitou em sua busca.

O usuário não pode simplesmente requisitar uma informação, que pode estar expressa através de diferentes seqüências de palavras.

Entre o que o usuário deseja e o que ele requisita, portanto, há uma série de estratégias que ele deve utilizar para melhorar os acertos nos resultados apresentados pelo buscador. O usuário tem de prever as palavras que não podem faltar aos documentos que apresentem a idéia buscada, sendo que ainda deve evitar, se possível, as palavras que possam ser comuns a textos que não apresentem a informação desejada.

Os principais problemas desse padrão de busca são: i) o usuário pode perder muito tempo tentando prever as palavras que compõem os documentos que satisfaçam sua busca, e ii) podem ser retornados como resultados muitos documentos irrelevantes por apresentarem as mesmas palavras usadas pelo usuário em sua busca.

O que leva os buscadores ao primeiro problema é não relacionar sinônimos ou paráfrases, o que obriga o usuário a tentar uma série de formas diferentes para expressar a mesma idéia do que quer encontrar. O outro problema é causado pela

incapacidade dos buscadores de perceber os significados que estão sendo expressos tanto pelo usuário em seus requisitos de busca quanto pelos textos contidos nos sites.

A funcionalidade esperada para um sistema de recuperação da informação contida em textos seria recuperar informações, não palavras. Idealmente, o usuário poderia digitar o que espera encontrar, sem se preocupar em prever com exatidão a forma como essa idéia estará escrita nos documentos que poderiam servir como resultados e sem se preocupar com a possibilidade de que as palavras que ele usou em sua pesquisa possam aparecer em outros documentos sem expressar a idéia desejada. Daí a urgência para a criação de um mecanismo de busca orientado a idéia.

1.2 A busca orientada a idéia

Os índices de *precision* e de *recall*[1] de um mecanismo de busca em textos, sugere-se aqui, podem ser aumentados a partir do tratamento lingüístico dos documentos pesquisados e das requisições dos usuários.

A *precision* pode estar ligada aos significados das palavras. Os documentos que são encontrados pelo buscador e que não nos interessam são aqueles que apresentam as palavras usadas pela pesquisa, mas com significados diferentes daquele que queremos. Então, para melhorarmos os índices de *precision* das buscas, o buscador deveria ser capaz de definir qual é o significado de cada palavra nos documentos e nas requisições de busca.

O *recall* pode estar relacionado à paráfrase. Os documentos que têm as informações desejadas, mas que não aparecem listados como resultados pelo buscador, possuem as mesmas idéias que o usuário procura, mas expressas com palavras diferentes daquelas preenchidas como pedido de busca pelo usuário. Para aumentar o *recall*, a busca não deveria se limitar a procurar as palavras exatamente

[1] Dentre os métodos usados para avaliar um programa de computador que recupere informações, como é o caso dos mecanismos de busca, existem duas medidas de avaliação muito utilizadas, a *precision* e o *recall*. A *precision* se refere à quantidade de resultados satisfatórios apresentados pela busca em relação ao número total de resultados apresentados. O *recall* se refere à quantidade de resultados satisfatórios apresentados pela busca em relação ao total de documentos satisfatórios que deveriam ter sido apresentados pelo buscador.

como o usuário as digitou, mas deveria ser capaz de encontrar textos que expressem a mesma idéia que o usuário procura, independentemente da forma expressa.

Nesse novo padrão, os textos não deverão ser entendidos como seqüências de palavras, mas como conjuntos de idéias relacionadas entre si. E o mesmo ocorre com os pedidos de busca do usuário. Como o buscador deverá ser capaz de "compreender" a idéia buscada pelo usuário, seus pedidos não podem ser expressos por palavras soltas, sem nenhuma relação de sentido entre elas. A forma pela qual as palavras estabelecem uma relação de sentido entre si, como veremos mais adiante, é através da relação sintática. Portanto, para que os pedidos de busca também sejam entendidos como idéias relacionadas, eles deverão ser expressos como pequenos textos (uma frase, uma expressão etc.). E a busca deverá se concentrar em encontrar os documentos que contiverem as idéias pedidas, de tal forma que seja uma busca por um pequeno conjunto de idéias (os pedidos), dentro de um conjunto maior de idéias (os arquivos de texto).

A partir de agora, será apresentada uma proposta de modelo de representação de textos para a utilização em buscadores orientados a idéia. Este modelo, ainda embrionário, visa reduzir um texto a um conjunto de idéias tratáveis computacionalmente.

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:

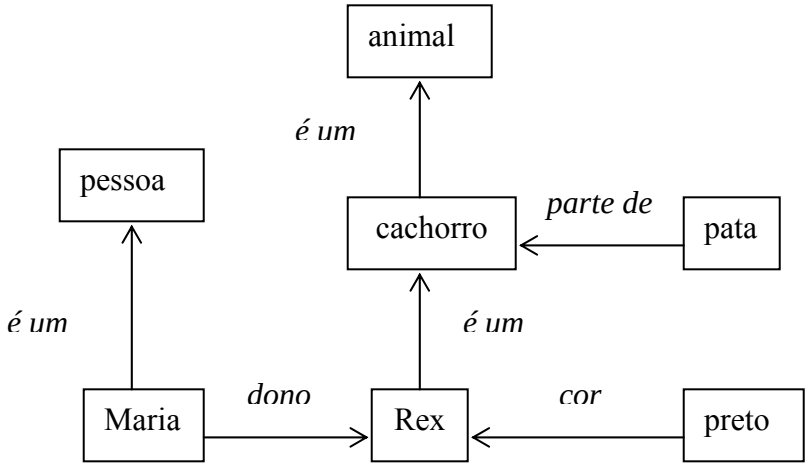


Figura 1 – rede semântica

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*.

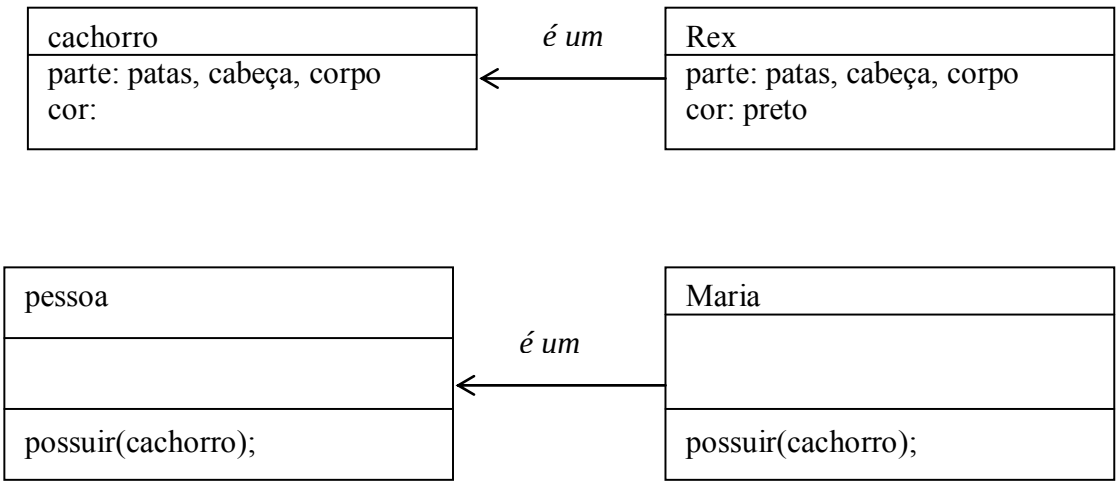


Figura 2 – modelo de *frames*

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.

3
A proposta

Visando desenvolver esse novo paradigma de busca, foi necessário fazer com que o buscador resolvesse os dois problemas levantados: i) definir os significados dos textos, e ii) reconhecer as paráfrases.

Direcionou-se a resolução dos dois problemas para seu tratamento como um só, unindo as duas idéias: distinção de significados e paráfrase. Dessa maneira, o significado de um texto será definido a partir do modo como ele possa ser reescrito. Definir o significado de uma palavra, expressão ou texto, portanto, significa, para nossa aplicação, saber tão somente como podemos parafraseá-los. E isso é suficiente (embora não necessariamente fácil, ou mesmo factível em grandes corpora). Como se trata de um modelo preliminar, assumimos a premissa de que a proposta é factível.

Do ponto de vista computacional, serão acrescentados dados semânticos aos dados já contidos nos documentos com formato de arquivos de texto. Assim, juntamente com as informações a respeito dos caracteres alfanuméricos que formam as palavras e das informações sobre a formatação dos textos[1], um arquivo de texto teria também as informações sobre os significados depreendidos no texto. É sobre essas informações semânticas[2] que o sistema de busca realizará suas pesquisas.

Ao acrescentar informação semântica aos dados de um arquivo de texto, pretende-se que o computador seja capaz de manipular o texto semanticamente.

Isso significa dar à computação o poder de trabalhar com a distinção de significados, isso é, compreender que uma determinada palavra está exercendo um determinado significado, e não outro, no contexto em que ela aparece. Como foi dito,

[1] Por exemplo, os arquivos de texto com extensão *.txt* representam os caracteres alfanuméricos através da codificação ASCII. Os arquivos de texto com extensão *.rtf* representam a formatação através de códigos escritos após a barra invertida (por exemplo, a quebra de linha é representada pelo código "\par") e o trecho de texto a qual recai sua aplicação é delimitado entre chaves ("{" e "}").

[2] A semântica nesta pesquisa é a parte da gramática responsável pelos significados, diferenciando-se da sintaxe, que estuda as regras de formação das sentenças, e diferenciando-se do componente textual da gramática, que estuda o relacionamento entre sentenças. Claro que essa é uma divisão didática já que, por exemplo, uma regra sintática, muitas vezes, é determinada por fatores semânticos. O adjetivo "semântico" significa, aqui, "relativo aos significados". As palavras "significado", "sentido", "idéias" e "acepção" são consideradas sinônimas na presente pesquisa.

determinar o significado de uma palavra (e mesmo de uma expressão ou texto) está relacionado à paráfrase.

A equiparação entre paráfrase e significado proporcionará algo valioso para que seja possível a busca orientada a idéia: está-se excluindo a idéia de referencialidade do sentido. Para definirmos, assim, qual significado uma palavra está exercendo num determinado contexto, não precisamos saber a que ela se refere, bastará sabermos com quais outras palavras ou expressões ela poderia ser reescrita. O significado de uma palavra é apenas esse centro para qual as possíveis paráfrases convergem. Isso quer dizer que, quando estivermos buscando um modo de representar as idéias de um texto, não nos importaremos com o que as palavras são usadas para se referirem, mas apenas com o processo de substituir palavras[3].

Em suma, não se pretende que o sistema computacional "entenda" o texto, mas que seja capaz de manipulá-lo como nós o fazemos a partir da nossa capacidade de produzir paráfrases, seja reescrevendo sentenças, seja compreendendo duas frases diferentes como tendo o mesmo conteúdo significativo, seja podendo resumir um texto, seja por sermos capazes de encontrar informações explicitamente escritas num texto, seja por podermos desambiguar uma sentença.

Se limitarmos a determinação de significado de um texto às suas paráfrases, então os processos de manipular a superfície semântica de um texto podem ser simplificados, por não precisarmos de grandes bases de conhecimento. A determinação de significados passa a ser um processo quase mecânico de levantar as formas como um texto pode ser reescrito. Sendo assim, exclui-se desses processos, em nossa proposta, todos as formas de raciocínio mais sofisticadas como as inferências, as conclusões, os preenchimentos de lacunas no texto, etc. Essas habilidades fariam parte não da superfície semântica da linguagem – já que, aqui, o significado se reduz à paráfrase –, mas de características próprias do conhecimento

[3] Existe um fato paradoxal nessa estratégia. Na tentativa de formular as bases lingüísticas de um buscador orientado a idéia, em que o usuário não se importaria com as palavras e, sim, com as idéias expressas por elas, foi definido que a mecânica por trás desse buscador não se importará com as "idéias" (entendida como "o que uma palavra se refere") e, sim, com o jogo de intercambiar palavras. Para fugir desse paradoxo, distinguiu-se "referencialidade" de "idéia". Nesta pesquisa, "idéia" é equiparada a "paráfrase" e não significa "referência". Note que para o usuário, essa distinção é transparente (no sentido que a informática usa), ele não se dará conta disso. Para ele, a busca será por informações e não por palavras escritas para representá-las.

do mundo, do conhecimento enciclopédico ou cultural, que, apesar de úteis para diversas aplicações, podem ser ignoradas para a busca orientada a idéia[4].

Aqui já temos nossa primeira base lingüística para conseguirmos fazer um buscador orientado a idéia: equipar significado a paráfrase.

3. 1
Representação semântica

Do exposto acima, é respondida a pergunta "como o computador pode buscar idéias e não palavras?" através da resposta: ao reconhecer paráfrases. Mas daí surge outra pergunta: como fazer para que o computador reconheça uma paráfrase?

É fundamental entendermos que caso se pretenda acrescentar dados semânticos aos arquivos de textos, então textos que compartilhem os mesmos significados também compartilharão certos dados semânticos.

Esse é o objetivo básico de se acrescentar informação semântica aos arquivos de textos: criar um modelo de representação das informações semânticas de um arquivo de texto, de tal forma que palavras e expressões que sejam paráfrases tenham obrigatoriamente que ser representadas da mesma forma. E, ao mesmo tempo, as palavras e expressões que não sejam paráfrases têm que ser representadas de maneiras diferentes.

E como representar os significados?

Se quisermos pensar em representar o significado de um texto, a maneira mais imediata que nos ocorre é partirmos dos significados das palavras que o compõem. Isso quer dizer que podemos pensar que o significado total de um texto é a soma dos significados de cada palavra que fazem parte dele[5].

[4] Alguns teóricos consideram a inconsistência, a implicação e outros conceitos que pressupõem a noção de verdade como fazendo parte dos estudos semânticos (Katz, 1978, 52-53). No entanto, como para nós semântica se liga à paráfrase, devemos excluir a análise da verdade de nossa particular semântica, sendo necessário que seja tratada por conhecimentos não exclusivamente lingüísticos como a Lógica. O único fenômeno que poderia ter algum resquício de um teste sobre a verdade e que entrará nesse modelo de representação é a contradição, ou, mais exatamente, a negação.

[5] Essa afirmação precisa será relativizada e contextualizada, o que faremos a seguir.

Devemos privilegiar o que há de intuitivo na estratégia de relacionar significado do texto ao significado de suas palavras. Certamente o significado de um texto parte do significado das palavras que o compõem, embora não esteja restrito a isso.

Entretanto, há um problema em considerarmos que o significado de um texto é dependente do significado de suas palavras. O significado de uma palavra no texto também depende do significado total do texto.

Como definir o significado de uma palavra, se a mesma palavra pode ter vários significados? A resposta que sempre ouvimos para essa pergunta é: pelo contexto. Contexto deve ser entendido como a relação que uma palavra estabelece com todo o texto.

Podemos concluir daí que existe uma íntima relação mantida entre significado da palavra e significado do texto. Isso porque o significado de um texto é depreendido a partir do significado de cada palavra que o compõe, ao mesmo tempo em que o significado de uma palavra é dado a partir da relação que esta estabelece com o restante do texto.

É possível partir, então, por dois caminhos a fim de encontrar uma representação para os significados de um texto: ou partir do texto e chegar às palavras ou partir das palavras e chegar ao texto.

Tentaremos partir primeiramente das palavras.

3. 1. 1
O significado do texto como o significado das palavras que o compõem

Como foi dito antes, o significado de um texto pode ser dado pelo relacionamento com o significado das palavras que o compõem. Mas o primeiro cuidado que se deve ter em relação aos significados das palavras é que, ao atribuir o significado a uma palavra para tentar buscar o significado total do texto, devemos nos restringir ao significado que esta palavra está exercendo naquele momento, sem nos preocuparmos com as sutilezas de seus sentidos.

Por exemplo, nas sentenças “O cão fugiu de casa” e “O cachorro fugiu de casa”, as palavras “cão” e “cachorro” devem ser perfeitamente entendidas nesses contextos como sinônimas. A elas deve ser atribuído o mesmo significado, mesmo que cada

uma das palavras tenha pequenas particularidades que as diferenciam. Se nunca se aceitar as palavras ou expressões como tendo o mesmo e exato sentido num contexto, nunca se poderá admitir que elas formem paráfrases, e, assim, o plano de representar da mesma forma as paráfrases será frustrado já em suas premissas.

O que importa, para a finalidade pretendida, é encontrar o significado que essas palavras têm em cada uso, portanto e da mesma maneira, em “Estou com uma fome de cachorro”, o substantivo “cachorro” não compartilha o significado que “cachorro” exerce em “O cachorro fugiu de casa”. Isso porque a palavra “cachorro” não pode ser parafraseada da mesma maneira nos dois contextos. Se elas não são parafraseadas da mesma forma, então têm sentidos diferentes e, dessa maneira, devem ser tratadas como palavras diferentes – como homonímias e não polissemias[6].

Concluindo: a noção intuitiva de representar o significado de uma frase relacionando-a aos significados das palavras que a compõem somente terá valor se limitarmos os significados dessas palavras ao contexto.

E mais, para cada uso de uma palavra (se este determinar paráfrases distintas), ela será tratada como se fosse uma palavra diferente.

Voltemos à questão de representar o significado de um texto através da soma dos significados das palavras que o compõem. Se atribuirmos uma letra a cada palavra, representando seu significado na sentença “João matou José”, como na ilustração

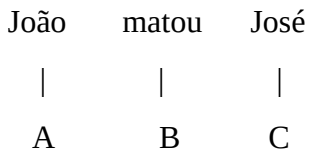


Figura 1 – união dos significados

então o significado da sentença seria dado pelo conjunto {A, B,C}.

[6] Como estamos apresentando um modelo particular de representação lingüística para ser usado em buscadores, cabe lembrar que é tão somente neste modelo que os dois significados de “cachorro” se apresentam como homonímiaa e não polissemiaa, pois, nesse modelo, “cachorro” deverá corresponder, segundo o exemplo, a duas entradas lexicais . E esse fato é que, para nós, corresponde à homonímia e não outros critérios tradicionais como o grau de distinção de significado.

Mesmo admitindo que o significado da sentença se relacione com o significado das palavras que a compõem, não podemos aceitar que esse relacionamento seja dado pela união ou soma dos significados de cada palavra.

Veja-se que, se tivermos uma sentença como “José matou João”, teríamos que representar com as mesmas letras cada uma das palavras, já que individualmente cada uma das palavras nos dois contextos tem seu significado inalterado.

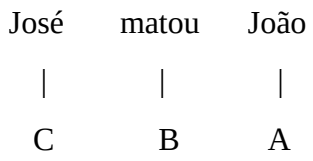


Figura 2 – seqüência dos significados

O conjunto que representaria o significado dessa sentença também seria {A,B,C}. Mas as duas sentenças não são paráfrases, não têm o mesmo significado e, portanto, não podem compartilhar da mesma estrutura que as represente.

A solução, também natural, para essa questão seria representar os significados de uma sentença não como o conjunto dos significados das palavras que a compõe, mas como uma seqüência ordenada desses significados.

Assim, no exemplo anterior, a sentença “João matou José” seria representada pela seqüência ABC, enquanto a sentença “José matou João” seria representada pela seqüência CBA, o que parece satisfazer nossa proposta.

Mas, quando nos deparamos com uma sentença como “o assassinato de José por João”, temos um problema. Temos que admitir que essa sentença é uma paráfrase de “João matou José” e, portanto, “assassinato” e “matar” devem ser representados pelo mesmo conceito.

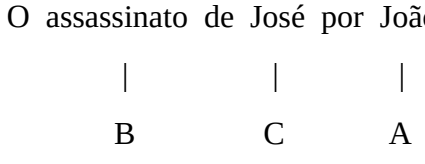


Figura 3 – seqüência dos significados em uma paráfrase

Dessa maneira, duas paráfrases estariam sendo representadas de forma diferente: a seqüência ABC para “João matou José” e a seqüência BCA para “O assassinato de José por João”, o que contradiz nossa proposta.

Como resolver isso?

Essa pergunta será deixada em aberto por ora, para tentarmos uma outra abordagem a fim de que consigamos uma forma de representarmos os significados do texto. Partiremos agora do texto para chegarmos no significado das palavras.

Como foi dito, o significado de uma palavra é definido no contexto. Contexto é uma expressão quase mágica, a que boa parte dos modelos teóricos recorre quando as questões de difícil sistematização surgem.

Temos que aperfeiçoar a idéia de contexto, portanto.

Como a palavra se relaciona com o texto para que seu sentido possa ser definido? O texto é composto de palavras, então o relacionamento de uma palavra se dá com as outras palavras do texto. E como essas palavras estão relacionadas? As palavras se relacionam através da gramática.

Torna-se então absolutamente necessário apresentar nesse momento o modelo gramatical em que me baseio para conseguir representar os dados semânticos de um arquivo de texto. Novamente, esse modelo é uma representação da linguagem para um fim prático já determinado, ele não pretende ser um modelo teórico que explique toda a realidade lingüística.

3. 1. 2
Considerações gramaticais

Por serem baseadas na semântica, ou melhor, na referencialidade, as definições que a gramática tradicional faz de classe de palavra – assim como o faz das funções sintáticas – são freqüentemente alvos de críticas. Por exemplo, é enormemente criticada a definição de substantivo como a palavra que dá nome às coisas, porque nem todo substantivo serviria para dar nome a alguma coisa.

Apesar desse ser um argumento bastante sensato, é quase impraticável tentar definir algumas classes gramaticais básicas sem recorrer à intuição, por isso é impossível fugir da definição pela referencialidade[7].

Isso se dá porque a função mais primária da língua é servir para organizarmos nosso pensamento e comunicá-lo aos outros. Percebemos o mundo através de nossos sentidos; é através do filtro da percepção que compreendemos as coisas e, portanto, filtramos também a língua pela percepção.

Do mundo que nos cerca percebemos objetos, pessoas, formas. As palavras que usamos para nos referirmos a isso são os nomes[8].

A classe de palavra mais básica que existe, para nós, é o nome. Nome, nesta proposta de modelo, não abarca o que a gramática tradicional conceitua como substantivo, mas tão somente as palavras usadas para nos referirmos às entidades[9] do mundo que podemos captar por algum de nossos sentidos. Para nós, são nomes: "bola", "cadeira", "fogão", "gato", "pessoa".

Também conseguimos perceber no mundo sensível que algumas características podem ser compartilhadas por essas entidades. Vemos plantas verdes assim como vemos um livro verde, por exemplo. A classe de palavra que se refere às propriedades que podem ser compartilhadas pelas entidades é o qualificador. São qualificadores: "pequeno", "baixo", "azul".

Esse mesmo compartilhamento que existe entre entidades com relação às propriedades também existe com relação às ações. Uma mesma ação pode ser executada por diferentes entidades; por exemplo: um copo cai, uma folha cai, uma pessoa cai.

[7] Como foi dito anteriormente, tinha sido excluída a referencialidade da noção de significado, e, portanto, da semântica. No entanto, ela terá sua importância para a sintaxe. E como iremos atribuir o significado das palavras através da sintaxe, por fim, a referencialidade influenciará, ainda que indiretamente, a semântica também.

[8] O conceito nome usado nesse modelo não é o mesmo o conceito que se faz normalmente de substantivo e nem mesmo de nome. Veremos.

[9] É preciso ressaltar que está sendo feito uso de conceitos diferentes para fazer referência às entidades e aos nomes. Entidades são as coisas que existem concretamente no mundo: o papel, a árvore etc. Os nomes são as classes de palavra que usamos para nos referirmos às entidades nessa particular realidade que é a língua. São nomes as palavras: "papel", "árvore". Para distinguir esses dois conceitos os nomes sempre virão entre aspas.

A classe de palavra que representa ações é o evento[10].

Usamos classes de palavra diferentes para nos referirmos a cada um desses aspectos encontrados na realidade. São categorias distintas, então possuem comportamentos distintos. Por exemplo, um nome obedece a padrões de flexão como o número, e têm algumas propriedades como o gênero. Cada classe de palavras obedece, portanto, a um padrão de terminação e/ou flexão.

É muito importante que chamemos a atenção para algumas características dessas classes. As entidades podem ser percebidas na realidade por si só, sem que necessariamente possuam essa ou aquela propriedade, sem que pratiquem essa ou aquela ação.

O mesmo não ocorre com as propriedades e as ações. Na realidade que nos cerca, as propriedades e as ações não existem a não ser como propriedades de alguma coisa ou ações que envolvam alguma coisa.

Uma propriedade como uma cor não se realiza a não ser como a cor de algo. Quando abrimos os olhos, não nos deparamos simplesmente com o branco, vemos uma parede branca, um guardanapo branco, uma caneca branca.

Da mesma forma, não podemos tirar uma foto de uma corrida, mas apenas fotografamos carros correndo, pessoas correndo, animais correndo.

Propriedades e ações só existem na realidade quando realizadas por entidades concretas do mundo. E o mesmo acontece com as palavras que usamos para representá-las. Qualificadores e eventos necessitam de nomes para serem realizados. Essa necessidade de complementação, quando transposta para a língua, é o que chamamos de uma relação argumentativa.

Como já se deve ter notado, foram definidas até agora classes que podemos reconhecer através da nomenclatura tradicional: os nomes seriam substantivos, os qualificadores seriam adjetivos e os eventos seriam verbos. No entanto, pelas definições feitas aqui para cada uma dessas classes, nem todo substantivo (como o conhecemos) poderia ser considerado um nome, nem todo adjetivo poderia ser considerado um qualificador, nem todo verbo poderia ser considerado um evento.

[10] A mesma distinção que foi feita entre os conceitos entidade e nome é feita entre os conceitos propriedade e qualificador e entre ação e evento. Por se tratarem de palavras, os qualificadores e os eventos serão grafados entre aspas.

A língua consegue nos proporcionar grande poder de abstração. Isso porque a língua é uma realidade própria e flexível. Do ponto de vista argumental, podemos definir nome como aquela palavra que serve de argumento para qualificadores e eventos. Outras palavras que não sejam nomes podem ocupar a posição de um nome, isso é, podem servir como argumentos de um qualificador ou de um evento. Para tanto, essa palavra deve receber certas terminações e variar conforme um nome.

Por exemplo, para que o evento "cair" ocupe a posição de nome, isso é, para que o evento "cair" possa ser complemento de um evento ou de um qualificador, ele tem que sofrer alterações morfológicas e, como "queda", pode então se posicionar onde um nome se posicionaria.

Uma palavra como "queda", no entanto, não pode ser considerada, em nosso modelo, um nome. Isso porque além de "queda" não ser a designação de uma entidade observada no mundo, ela apresenta uma carência de argumentos, o que contradiz nossa definição de nome. Mesmo ocupando a posição de um nome, "queda" carrega consigo a mesma necessidade de um argumento que possuía enquanto evento.

Palavras que podem ocupar a posição de nomes recebem a denominação de substantivos no nosso modelo. A palavra "queda", portanto, é um tipo de substantivo. Substantivos que têm o significado de um evento são chamados em nosso modelo de substantivos-evento.

Como um nome obviamente pode ocupar o lugar de um nome, ele também é um substantivo. Por isso, substantivo é uma superclasse em que nomes, substantivos-evento e outros estão contidos[11].

Essas subdivisões acontecem também com verbos (que incluem eventos e outras palavras que ocupam a mesma posição que pode ocupar um evento) e com adjetivos (que incluem qualificadores e outras palavras que ocupam a mesma posição que pode ocupar um qualificador).

[11] No presente modelo, a pesar de termos nomeado como substantivo a superclasse que abarca nomes e outras palavras que ocupam a posição de nomes, não estamos nos referindo a outras conceituações que substantivo possa ter eventualmente em outras teorias e autores. O conceito substantivo, aqui, é um conceito próprio a este modelo particular de representação da língua. Quando nos referirmos a substantivo, portanto, estaremos nos referindo a esta conceituação particular. Sempre que quisermos mencionar outras conceituações para a palavra "substantivo", isso será explicitamente referido.

A relação argumentativa é fundamental à gramática de uma língua. E, em nosso modelo, é a base da sintaxe. De tal modo que, para os intuitos dessa pesquisa, toda e qualquer função sintática pode ser compreendida como uma relação entre argumentos.

Para nós, uma função sintática existe quando uma palavra rege outra, isto é, quando uma palavra precisa de outra para a completar[12]. Sempre que existir a necessidade de uma complementação, existirá uma relação de argumento.

Isso fica claro com relação às palavras lexicais como verbos, substantivos e adjetivos (quer dizer, podemos relacionar a função sintática estabelecida entre um verbo e um substantivo com a relação argumental existente entre as duas palavras), mas isso também ocorre com as palavras gramaticais como preposições, conjunções e pronomes?

Vejamos o caso dos eventos. Alguns eventos pedem algo mais que nomes. O evento "bater", como sabemos, pede dois complementos; no entanto, um dos nomes pedidos deve ser intermediado pela preposição "em". A preposição, apesar de obrigatória, pode ser vista por nós como um acidente. O real argumento continua sendo apenas a palavra que corresponde à classe gramatical selecionada pelo evento "bater", que, no caso, é o nome. Entre os argumentos que pedem a intermediação de outras palavras e os que não a pedem não existe real diferença, a não ser a mediação casual de uma ou mais (como veremos) palavras.

No entanto, a presença da preposição, mesmo que ela não seja exatamente o argumento do evento, é justificada pela relação argumentativa estabelecida entre evento e nome.

Podemos aplicar esse mesmo processo as demais palavras gramaticais. Todas as palavras gramaticais (artigos, pronomes, conjunções etc.) também serão justificadas, em nosso modelo, por relações argumentais, como veremos mais à frente.

O que é necessário por ora é saber que as palavras-chave para compreendermos o relacionamento de uma palavra com as demais num texto são "relação argumentativa".

[12] A equiparação entre função sintática e regência me foi ensinada pelo grande mestre José Carlos Azeredo. Aqui foi apenas acrescentada a equiparação com a relação argumentativa.

A relação argumentativa é a relação gramatical por excelência. Ela organiza a disposição das palavras na frase e é ela ainda que gerencia o uso das palavras gramaticais como preposições e artigos.

E o ponto mais importante para nós é que uma relação argumentativa é uma seleção. Quando se afirma que uma palavra pede argumentos, também se afirma que essa mesma palavra seleciona certas particularidades de seu argumento, determinando assim que apenas algumas palavras possam compor sua relação argumentativa.

Tais particularidades selecionadas por uma relação argumentativa são de duas naturezas: classe de palavra e semântica.

A primeira seleção que uma palavra faz de seu argumento diz respeito à sua classe de palavra. Por exemplo, uma propriedade como “alto” necessita de uma palavra sobre a qual irá incidir, tendo tal palavra que pertencer à classe dos nomes.

Mas não é a qualquer nome que se poderá atribuir a propriedade de ser alto. É difícil imaginar um cheiro alto. E, em outros momentos, dependendo das características semânticas do complemento, a palavra que o exige tenderá a ter um significado ou outro. Por exemplo, um muro alto é aquele muro com extensão vertical elevada, enquanto uma nuvem alta é uma nuvem postada a grande elevação com relação ao solo, e, ainda, um som alto é um som com intensidade elevada.

Se existem diferenças de sentido, e, portanto, paráfrases distintas, então teremos palavras diferentes em nosso modelo. Nesse exemplo rápido, conseguimos identificar pelo menos três homônimos de “alto”. Cada um seleciona semanticamente o nome que lhe servirá de complemento de forma distinta[13].

Concluindo, palavras que pedem argumentos selecionam a classe de palavra a que esse argumento pertencerá e selecionam características semânticas que esses argumentos deverão ter.

[13] Embora possa parecer inusual, nesta pesquisa sempre é o adjetivo que rege o nome como seu complemento e, portanto, o adjetivo seleciona as propriedades semânticas do nome. Como foi explicado anteriormente, assim como ocorre com os eventos, a necessidade de argumentos de um qualificador (e, por conseqüência, de um adjetivo) advém de uma necessidade que uma propriedade tem de se realizar sob a forma de uma entidade que possua tal propriedade. Essa foi a justificativa encontrada para definir essa relação argumentativa, então somente este poderá ser o sentido dessa seleção semântica: o adjetivo seleciona o substantivo, e não o contrário.

Isso já é suficiente para nós, por enquanto. Podemos voltar agora para o que nos interessa imediatamente, ou seja, a representação dos significados das palavras de um texto a partir do contexto.

3. 1. 3
O significado da palavra definido a partir do significado do texto

Uma palavra pode ter uma série de significados. Foi dito anteriormente que o significado pode ser determinado no contexto. Isso quer dizer que o significado de uma palavra pode ser determinado pela relação entre as palavras do texto. Essa relação é a relação argumentativa.

Tomemos como exemplo a sentença “o macaco correu”. Podemos pensar em pelo menos dois sentidos para a palavra “macaco”: o animal e a ferramenta. Para a palavra “correr”, também podemos pensar em dois sentidos: um sentido seria “andar rápido”; para tanto, o complemento dessa ação terá que ser um ser vivo. Outro sentido seria o de “escoar” e, para tanto, o complemento dessa ação deverá ser um líquido.

Cotejando as quatro possibilidades de sentido, podemos encontrar o significado de cada uma das palavras porque existe apenas uma combinação (ou unificação) em que a relação de argumento entre as duas palavras juntas é satisfeita.

Assim fica claro perceber que podemos definir o sentido de uma palavra a partir das relações argumentais de que ela participa, seja por pedir um complemento, seja por servir de complemento.

Ainda assim não conseguimos satisfazer nosso propósito de representar os significados. Isso porque os significados que demos para ambas as palavras também são compostos por palavras que podem possuir vários significados. Temos que encontrar uma forma mais precisa e menos ambígua para representar os significados das palavras.

Será repetida aquela estratégia utilizada quando da análise do significado do texto como composto pelo significado das palavras e atribuir para cada sentido uma letra. Sendo assim, a palavra “macaco” pode ter um significado A (um animal) ou um significado B (uma ferramenta). A palavra “correr” pode ter um significado C (andar

rápido) ou um significado D (escorrer). Quando “correr” tem o significado C, pede como complemento uma palavra que tenha o significado E (ser vivo) e quando tem o significado D, pede como complemento uma palavra com o significado F (líquido).

Desta maneira, não conseguimos mais unificar os significados. Na verdade, antes conseguíamos unificar esses significados apenas porque raciocinávamos a partir da premissa de que todo animal é um ser vivo, como se o significado de “ser vivo” estivesse contido no significado de “animal”.

Ainda não conseguimos resolver nosso problema. No entanto, tivemos uma percepção valiosa: algumas idéias são idéias complexas, no sentido de que são idéias compostas por outras idéias especificadas. Um animal é um ser vivo com algumas características particulares que o especificam dentre os demais seres vivos.

Podemos aproveitar esse raciocínio e voltarmos para o nosso caminho anterior, em que definíamos o significado do texto a partir do significado das palavras que o compõem.

Aproveitando nosso antigo exemplo “João matou José”, o verbo “matar” pode ser decomposto em “causar a morte de”. Vamos, então, reconceituar as palavras:

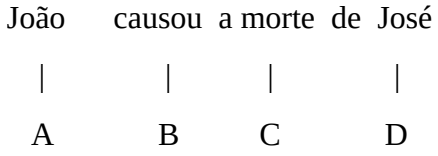


Figura 4 – decompondo um significado

Dessa maneira, o verbo “matou” em “João matou José” não terá apenas um conceito a ele relacionado, mas deverá ter tanto o conceito de “causou” como de “morte” para continuar compondo paráfrase. E, mais do que isso, esse exemplo é bem claro para mostrar que entre “causou” e “morte” é estabelecida uma relação de especificação. Isso porque é somente quando alguém causa o evento específico da morte de outra pessoa que podemos dizer que esse alguém matou essa outra pessoa.

A relação entre palavras, como já se disse antes, é uma relação argumental. Vemos nesse exemplo que a relação argumental é análoga à relação de especificação entre os significados.

Estamos chegando assim mais perto de onde queremos. Vamos representar a relação argumental através de uma seta que parte da palavra regente e aponta para o seu argumento. Dessa forma, teremos:

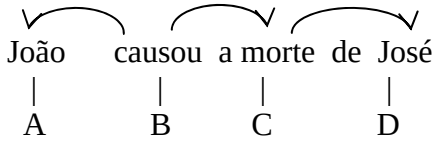


Figura 5 – relação argumental

Como foi dito, uma palavra ao servir de argumento estará especificando o sentido da palavra regente. Desse modo, podemos equiparar a relação argumentativa com a relação de especificação que percebemos que as idéia estabelecem entre si. Desse modo, vamos atribuir setas também aos significados destas palavras. Assim, se há uma seta partindo da palavra "causou" e apontando para a palavra "João", então uma seta partirá da letra que representa o significado de "causou" e apontará para a letra correspondente a "João". Desse modo, teremos:

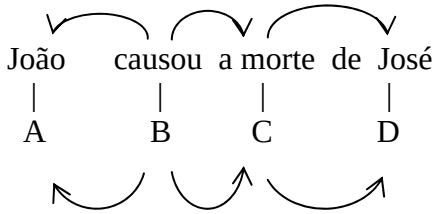


Figura 6 – significados

Esse esquema representa as relações argumentativas entre palavras e as relações de especificação entre significados.

Voltando à sentença "João matou José", o verbo "matou" é composto tanto pela idéia de "causou" quanto de "morte", já que é parafraseada por "causou a morte".

Dessa maneira, “matou” terá de ser representado não apenas por uma letra, mas tanto pela representação de “causou” quanto pela de “morte”.

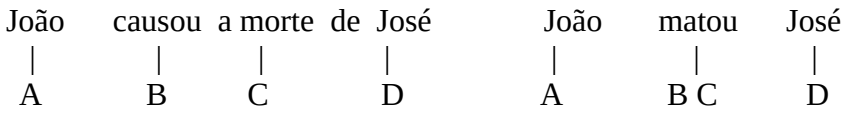


Figura 7 – redistribuindo os significados

Se formos representar as relações argumentais de “João matou José” por setas, e por analogia, atribuirmos as mesmas setas às letras que correspondem a seus significados, teremos um problema. O verbo “matou” não é composto apenas por uma letra, mas por duas, já que ele pode ser parafraseado pela expressão “causou a morte”. A qual das duas letras a seta que representa uma relação argumental de “matou” se ligará?

Ora, se estamos pretendendo representar paráfrases da mesma maneira, então o esquema de setas ligando as letras correspondentes ao significados das palavras de ambas as frases devem ser idênticas. Sendo assim, teremos:

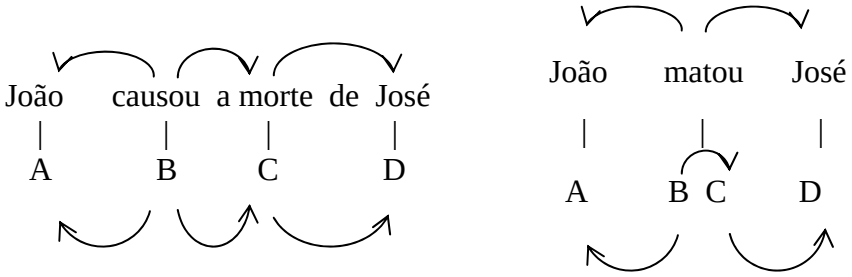


Figura 8 – atribuindo setas iguais às paráfrases

Podemos olhar a estrutura dos significados, representada por setas e letras, como um grafo. A teoria dos grafos é um modelo matemático muito utilizado pela Computação. Um grafo é usado para representar de forma gráfica alguns fenômenos. Um grafo é composto por elementos e traços que representam alguma relação estabelecida entre esses elementos.

Por exemplo, vamos imaginar que o conjunto dos elementos representados por um grafo seja composto pelo nome de alguns animais.

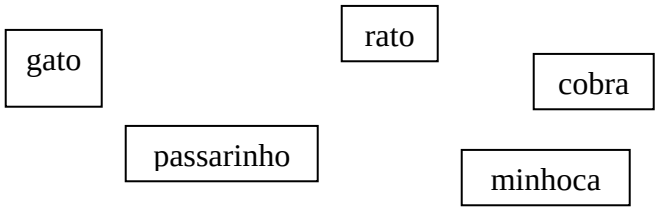


Figura 9 – nomes de animais

Vamos supor que nosso grafo represente a relação “caça”. Dessa maneira se um animal for caçado por outro, devemos desenhar um traço unindo os dois animais. Por exemplo, o passarinho é caçado pelo gato, portanto, representamos essa relação assim:

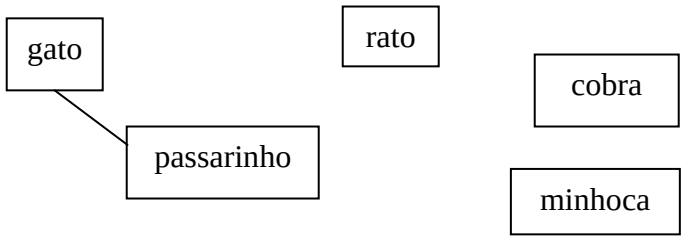


Figura 10 – a relação caça

Se representarmos por um traço, ligando cada animal caçado a seu respectivo caçador, nosso grafo ficará assim:

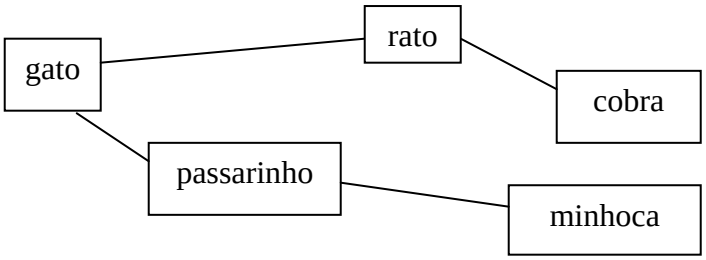


Figura 11 – grafo da caça

Um grafo é composto por elementos, no caso os animais, e aos elementos de um grafo damos o nome de nós. Um grafo também é composto por traços, no caso a relação “caça”, às relações estabelecidas entre os nós damos o nome de arcos de um grafo.

Portanto, um grafo representa uma situação que possa ser decomposta em elementos e em relações estabelecidas entre esses elementos. Note que, para poder ser representado por um grafo, cada relação sempre se dará entre dois elementos.

Os matemáticos notaram que todo grafo tem uma série de propriedades, independente de ele representar essa ou aquela realidade.

Por exemplo, vamos desenhar um novo grafo. Nele a ordem como aparecem os animais é diferente.

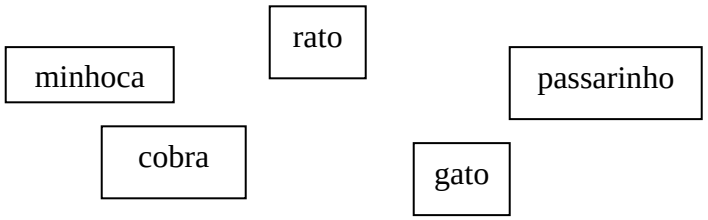


Figura 12 – grafo com outra configuração

Mas, como estamos usando esse novo grafo também para representar a mesma situação, isto é, a caça entre esses animais, então as relações entre os animais devem permanecer as mesmas. Isso quer dizer que se no nosso primeiro grafo ligamos “passarinho” a “gato” porque o gato caça o passarinho, no nosso novo grafo, também ligaremos “passarinho” a “gato”. Então o novo grafo fica assim:

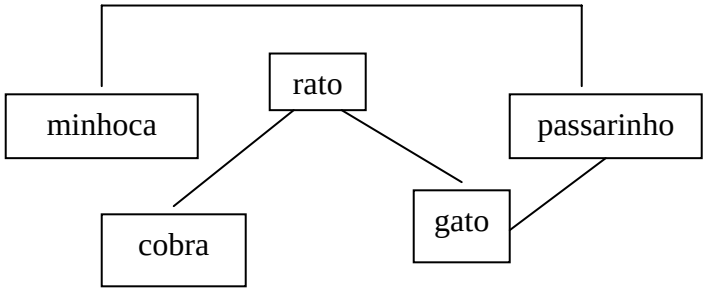


Figura 13 – novo grafo de caça

Apesar de terem configurações diferentes, os dois grafos estariam sendo usados para representar a mesma realidade. Portanto, dizemos que esses grafos são congruentes. Os matemáticos perceberam que, independente de qual seja a realidade representada por dois grafos, há como se saber se dois grafos são congruentes. Para dois grafos serem congruentes basta que eles sejam compostos pelos mesmos elementos e que as relações entre os elementos sejam as mesmas.

Vamos olhar novamente aquelas estruturas de setas e letras que tínhamos usado para representar os significados das sentenças "João causou a morte de José" e "João matou José". Se olharmos somente para as estruturas de setas e letras como um grafo, perceberemos que as duas estruturas são congruentes, porque têm os mesmos elementos e as mesmas relações entre esses elementos.

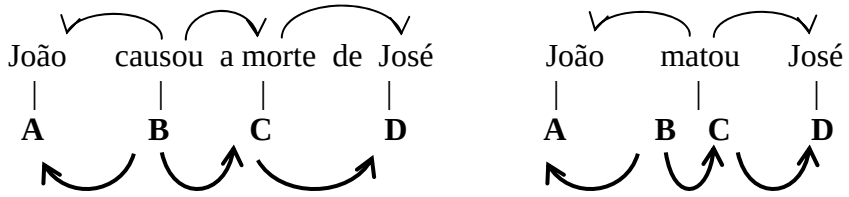


Figura 14 – grafos congruentes

Nosso grafo é um tipo particular de grafo, em que o arco (a relação) é orientado, isso é, é representado por uma seta[14]. A seta faz toda diferença, pois uma seta que parte de um elemento *A* e aponta para um elemento *B* não representa a mesma relação de uma seta que parte de *B* e aponta para *A*. Assim, apesar de terem os mesmos elementos estes dois grafos não são congruentes:

Figura 15 – exemplo de grafos não congruentes

[14] A relação que deve ser representada por uma seta num grafo é, na verdade, uma função, isso é, uma relação entre dois termos em que há uma direção – um elemento está em função de outro. As relações argumentais e as relações de especificação são, por isso, funções. Mas não deixam de ser relações, afinal, apesar de nem toda relação ser uma função, toda função é uma relação.

Voltando às nossas frases, se admitimos que os dois grafos são congruentes, então agora as paráfrases terão exatamente a mesma representação.

A primeira etapa de nosso objetivo principal foi alcançada, já que pudemos representar duas paráfrases da mesma forma. Mas será que conseguimos a segunda parte de nosso objetivo, isso é, será que conseguimos representar de forma diferente duas frases escritas com as mesmas palavras, mas que não sejam paráfrases?

Vamos desenhar uma estrutura de letras e setas para uma outra sentença, em que os sujeitos sejam alterados. As relações sintáticas estabelecidas entre as palavras mudam, afinal temos outro sujeito e outro objeto. Então as relações entre os significados das palavras também mudam[15]. Se as relações mudam, mudam também as representações. Vejamos:

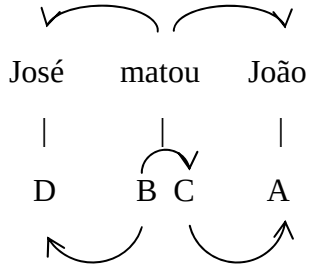


Figura 16 – grafo de uma sentença não paráfrase

Note-se que os grafos que representam os significados de “João matou José” e de “José matou João” não são congruentes, apesar de possuírem os mesmos elementos. Vejamos que, em “João matou José”, o elemento *B* está apontando para *A*, enquanto, em “José matou João”, o elemento *B* está apontando para *D*. O elemento *C* também aponta para elementos diferentes nas duas sentenças.

Com essa representação, conseguimos o nosso propósito: representamos os significados de tal forma que paráfrases possuem a mesma representação; enquanto frases com sentidos diferentes, mesmo que grafados com as mesmas palavras, possuem representações diferentes.

[15] Lembre-se que se há uma relação sintática entre duas palavras, isso significa que o significado de uma palavra está especificando o significado de outra palavra.

É defendido aqui que a melhor forma de tratar o significado de um texto a partir do significado de cada palavra é através de uma estrutura que se assemelhe a um grafo.

Para podermos representar os significados de um texto (uma sentença é um pequeno texto), são necessários dois alicerces básicos: i) a equivalência entre significado e paráfrase, e ii) a equivalência entre a relação sintática estabelecida entre duas palavras e a relação de especificação estabelecida entre duas idéias. Essas são as duas principais bases lingüísticas desenvolvidas pela presente pesquisa para a busca orientada a idéia.

Ao considerarmos que o significado de uma palavra num contexto é tão somente suas paráfrases, facilitamos a sua sistematização para um fim como a busca orientada à idéia. Para tanto, temos que admitir que existam paráfrases perfeitas e que os significados distintos de uma palavra determinam homonímias.

Percebemos que o significado de uma palavra nem sempre é uma idéia simples, isso é, um significado pode ser composto do significado de outras palavras. Muitas vezes, a idéia de uma palavra engloba o significado de uma expressão, isso é, o significado total de duas ou mais palavras. Percebemos também que, quando o significado de uma palavra é uma idéia complexa, isso é, uma idéia composta por mais de uma idéia simples, uma idéia especifica outra idéia. Dessa forma, o significado de uma palavra é composto por uma ou mais idéias. Mas o significado não é apenas o conjunto de uma ou mais idéias, é também dado pelo conjunto de relações de especificação estabelecidas entre suas idéias.

Notamos que, quando uma palavra serve de argumento para outra, esse complemento especifica o significado de sua palavra regente. Dessa forma, a relação sintática entre duas palavras pode ser vista por nós como uma relação de especificação entre as idéias de uma palavra e as idéias da outra palavra.

Representamos as relações de especificação entre as idéias como um grafo. A estrutura de grafos é uma idéia interessante porque, de uma só vez, representa o significado das palavras que compõem o texto e o significado do próprio texto.

Outra propriedade dos grafos levantada pela Matemática é que uma parte de um grafo pode ser olhada como um grafo. Dessa forma:

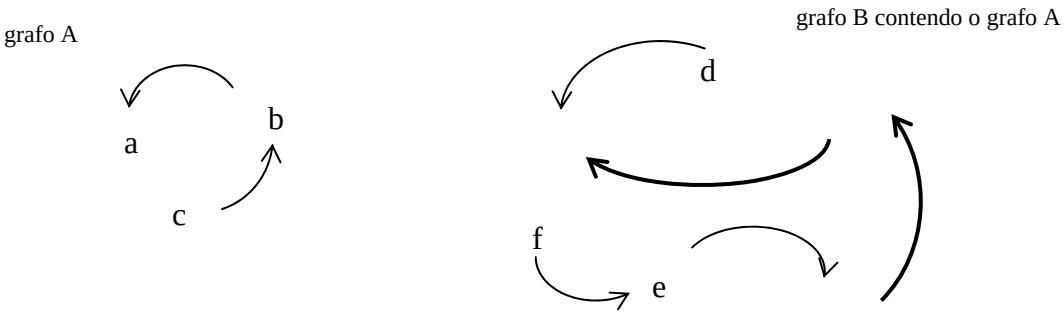


Figura 17 – um grafo contendo outro grafo

Isso quer dizer que um pequeno grafo pode ser encontrado como parte de um grafo maior. Quando um grafo pode estar contido em outro grafo, chamamos a ele de subgrafo.

Note-se que o grafo que representa individualmente o significado de uma palavra é um subgrafo do grafo que representa o significado de uma sentença. Veja com nosso exemplo:

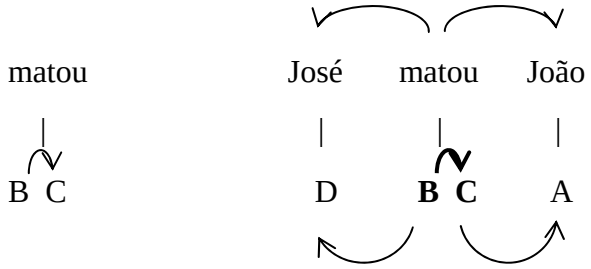


Figura 18 – subgrafo

Aos grafos que representam o significado de uma palavra ou de uma sentença chamamos em nosso modelo de estrutura de conceitos.

Através dessa estrutura de conceitos de toda uma frase é que podemos trabalhar com a habilidade que a língua nos proporciona de entendermos frases absolutamente diferentes como tendo o mesmo significado. Isso porque frases que têm o mesmo conteúdo significativo deverão ser reduzidas a uma mesma estrutura de conceitos, ou melhor, a estrutura de conceitos de cada uma das frases que sejam paráfrases deverá possuir os mesmos conceitos e as mesmas relações entre seus conceitos. O que as

distinguirá será a distribuição dos conceitos entre as diferentes palavras de cada frase, além das estruturas sintáticas.

Com relação à computação, trabalhar com grafos permitirá uma enorme flexibilidade nas buscas. Há décadas a Matemática e a Computação conhecem a habilidade de procurar um grafo como subgrafo de um grafo maior. Uma característica interessante para esta pesquisa é a de que uma procura por um grafo independe da seqüência entre os nós, da distância entre eles e de que haja outros nós os intercalando. O que importa é que no grafo analisado exista cada um dos nós buscados e cada uma das relações entre os nós correspondentes. Veja um exemplo:

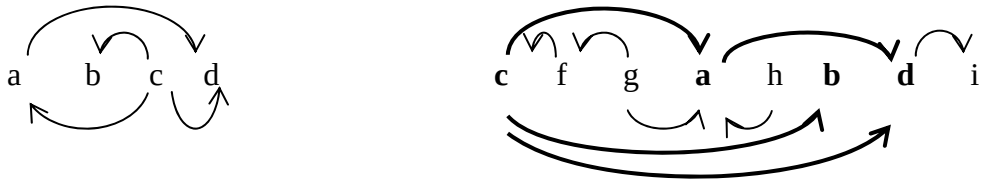


Figura 19 – habilidades de um grafo

Apropriando-nos dessa característica da teoria dos grafos para as estruturas de conceitos, ganharemos a capacidade de encontrar uma idéia que não apareça escrita seqüencialmente numa frase, mas que apareça decomposta em várias passagens do texto e consideramos que, ao final da leitura, a idéia desejada aparecerá completa.

Além disso, poderemos encontrar hipônimos se procurarmos por seus hiperônimos no buscador. Por exemplo, se procurarmos “animal”, serão apresentados textos em que apareça a palavra “gato”. Isto porque “gato” conterá em sua estrutura de conceitos toda a estrutura de conceitos correspondente a “animal”, já que faz parte da significação de gato ser um animal. Vejamos este exemplo de busca:

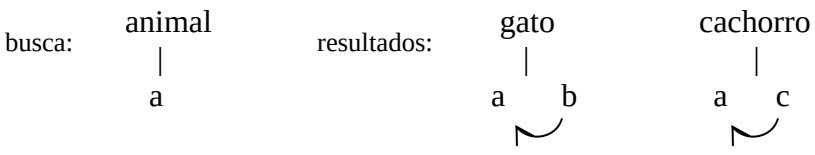


Figura 20 – grafo de hipônimos contendo o grafo de hiperônimos como subgrafo

Muitas outras habilidades lingüísticas poderão ser viabilizadas por esse modelo de representação lingüística, como a capacidade de reconhecer nuanças de significado ou de mostrar como uma idéia pode aparecer tanto sintetizada numa palavra quanto de forma analítica num pequeno texto.

É hora de formalizar melhor o modelo de representação dos significados, passando agora para a apresentação inicial de suas principais características.

3. 2
Formalizando a proposta

O modelo de representação gráfica da semântica existe para uma finalidade: a busca orientada a idéia.

Para conseguir fazer uma busca por idéias, e não por palavras, chegamos à conclusão de que uma boa estratégia seria acrescentar informação a respeito dos significados que são extraídos do texto. Essas informações semânticas é que serão pesquisadas durante as buscas orientadas a idéia e não as palavras escritas nos textos e nos requisitos dos usuários.

Também concluímos que o significado de um texto se relaciona com o significado das palavras que o compõem, embora o significado de um texto não seja a soma nem a seqüência dos significados das palavras que dele fazem parte.

O significado de cada palavra poderia ser representado, assim, como um conjunto de idéias (que daqui para frente chamaremos de conceitos) representadas por enquanto por letras. Esses conceitos também não são somas nem seqüências. Alguns conceitos especificam outros conceitos para tornar uma idéia singular entre tantas outras idéias. Então, escolhemos representar o significado de uma palavra por uma estrutura conceitual, em que alguns conceitos se relacionam dando especificidade uns aos outros. Essas estruturas conceituais podem ser representadas por um grafo orientado, cujos nós são os conceitos e as setas ligam dois conceitos em que haja uma relação de especificação, sempre apontando em direção à palavra que está especificando a outra.

As palavras que compõem o texto seriam substituídas pelas estruturas de conceitos que representem seus significados. Essas estruturas de conceitos seriam

unidas pelas relações sintáticas estabelecidas entre as palavras, já que uma relação sintática também é uma relação de especificação. Dessa forma, o significado final de um texto seria então uma grande estrutura de conceitos.

A representação da semântica é um recurso dentro de uma seqüência de outros passos que o computador executa para conseguir realizar a tarefa de buscar idéias.

Para formalizar esse modelo de representação, deve-se entender (ainda que genericamente) como será o caminho da execução pelo computador até a realização da busca por idéias, para, então, saber em que momento do processo da busca orientada a idéia esse modelo de representação entrará.

A busca orientada a idéia começa no momento em que o usuário digita seu requisito de busca e termina quando o computador retorna o *link* para os documentos que contenham a idéia do que foi requisitado. Tem-se, portanto, três grandes momentos ao computar esse processo:

Figura 21 – algoritmo 1

Essas etapas podem ser ainda subdivididas. A etapa “requisito de busca”, por exemplo, é composta de outros passos. O primeiro passo consiste basicamente de o programa esperar o usuário digitar o que quer buscar. A requisição do usuário deve ser transformada em uma estrutura de conceitos – afinal ela é o objeto da busca. Assim, “requisito de busca” pode ser dividido em:

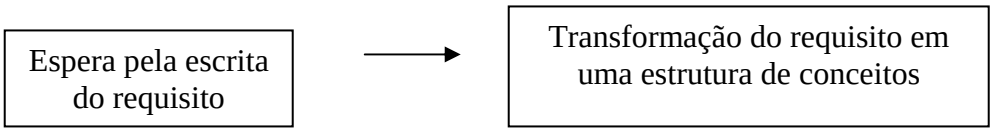


Figura 22 – algoritmo 2

A etapa “busca” consiste basicamente em encontrar o texto da base de dados que contenha a estrutura de conceitos requisitada pelo usuário. Portanto, os textos da base de dados deverão previamente ter sido também transformados em estruturas de conceitos.

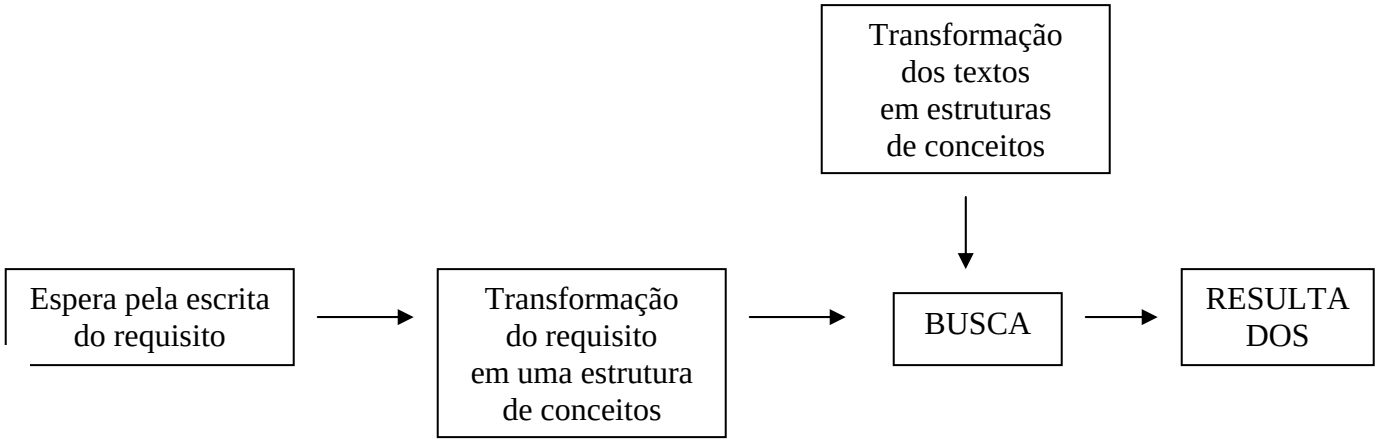


Figura 23 – algoritmo 3

Esses dois passos, “transformação do requisito em uma estrutura de conceitos” e “transformação dos textos em estruturas de conceitos”, é que nos interessam e, portanto, apenas eles serão decompostos aqui nesse momento.

Os dois passos consistem basicamente no mesmo processo, apenas agindo sobre dados diferentes: um passo transforma os requisitos escritos pelo usuário e o outro transforma os textos retirados da base de dados. Vamos tratá-los em conjunto, pois, dessa maneira, os passos em que subdividiremos um deles serão os mesmos em que se subdivide o outro.

Na transformação de um texto – seja ele de um requisito de usuário ou dos arquivos buscados – em estrutura de conceitos, os dados de entrada são compostos de uma seqüência de caracteres. Os dados de saída serão justamente o grafo da estrutura de conceitos.

O primeiro passo para essa conversão é reconhecer que essa seqüência de caracteres alfanuméricos é um conjunto de palavras. Passada essa etapa, o programa

atribuirá a cada palavra a estrutura de conceitos correspondente a seu significado. A estrutura de conceitos final será dada pela junção das estruturas de conceitos de cada palavra.

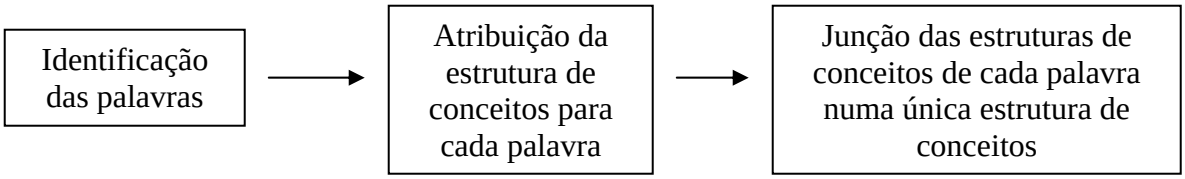


Figura 24 – algoritmo 4

Para atribuirmos uma estrutura de conceitos a uma palavra, devemos definir qual é o significado que essa palavra está exercendo no contexto. Como já se discutiu atrás, é a partir da análise sintática que desambiguamos essas palavras.

Então o passo "Atribuição da estrutura de conceitos para cada palavra" pode ser decomposto em outros passos. Primeiramente, consulta-se o dicionário eletrônico[16] para saber todos os significados que a palavra pode ter. No momento da consulta ao dicionário, o programa também deverá receber outras informações: as classes que a palavra pode ocupar em cada significado e quais argumentos ela pede em cada significado.

Feito isso com todas as palavras, o programa pode então checar qual combinação de classe de palavra, estrutura de conceitos e argumentos gera uma unificação possível.

Se não se encontrar nenhuma unificação possível, a sentença é agramatical. Se for encontrada mais de uma unificação possível, a sentença é ambígua. Deve-se sempre lembrar que todas as palavras devem participar, de alguma forma, de uma relação argumental, senão a sentença também será agramatical.

Vejamos como isso tudo ficará no nosso algoritmo esboçado:

[16] O dicionário eletrônico é um banco de dados em que as entradas são palavras e onde se atribui uma série de informações lingüísticas a essas palavras, de forma que um programa possa computar essas informações.

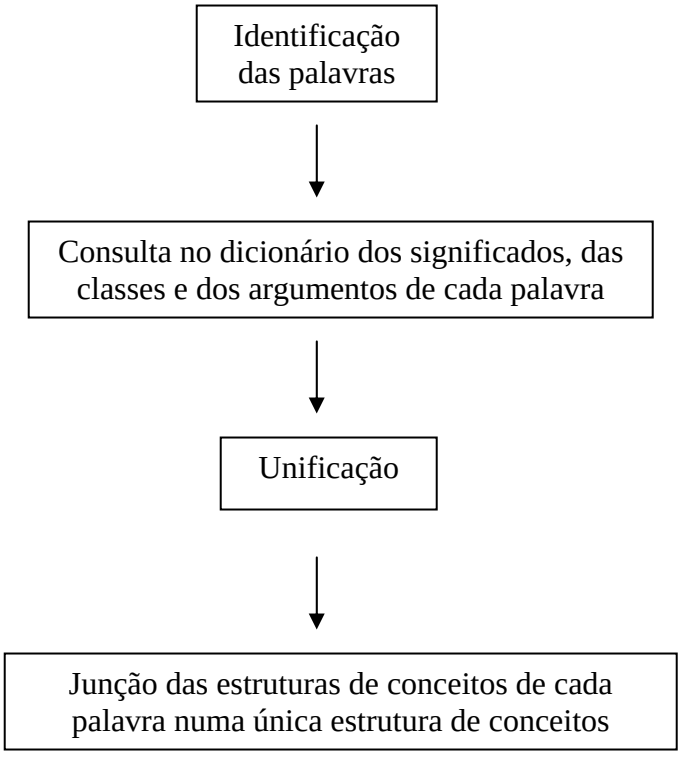


Figura 25 – algoritmo 5

Como podemos ver, é no dicionário eletrônico que aparecem tanto as informações sobre o significado de uma entrada (sua estrutura de conceitos), quanto as informações sobre a construção sintática que a palavra gera (sua relação argumental). Então é nas entradas do dicionário eletrônico que devemos formalizar as representações do significado e também as representações da sintaxe (já que estamos tratando toda relação sintática como relação argumental).

3. 2. 1
Formalizando as entradas lexicais

Como foi exposto, o presente modelo de representação da semântica será usado num momento particular do programa de busca orientada a idéia. Esse momento corresponde ao dicionário eletrônico.

Para cada entrada do dicionário eletrônico, devem ser informados a classe de palavra a que a entrada corresponde, o significado que essa palavra possui e os argumentos pedidos por essa palavra quando assume esse significado. Outra

informação importante, que iremos abordar em breve, diz respeito à disposição dessa palavra e de seus argumentos na sentença.

A informação sobre a classe de palavra, como a pequena amostra apresentada anteriormente já pôde evidenciar, obedece a uma classificação ligeiramente diferente das distribuições em classes propostas por outros modelos e autores. Isso porque estamos otimizando as classificações, bem como as regras sintáticas, para a nossa pretendida finalidade.

O significado da entrada lexical é informado exatamente através do grafo da estrutura de conceitos, isso é, o significado é a estrutura de conceitos. É esse grafo que deverá substituir a palavra na sentença para depois ser unido a todos os grafos correspondentes às estruturas de conceito das demais palavras do texto, formando assim o grande grafo que representa as idéias do texto.

Tanto classe de palavra quanto estrutura de conceitos serão tratados nos próximos tópicos desta exposição. Agora nos deteremos mais nas relações argumentais das entradas lexicais e na disposição das palavras na sentença.

Uma particularidade do modelo de representação semântica aqui proposto é que o conjunto de regras que formam a sintaxe não será apresentado fora do dicionário eletrônico. É nas entradas lexicais que toda e qualquer informação sintática aparece. Isso quer dizer que, para ser feita a análise sintática de uma sentença, todas as informações sobre a sintaxe serão adquiridas pelo programa através das informações contidas nas entradas do dicionário eletrônico correspondentes às palavras da sentença.

Isso se dá, porque equiparamos a sintaxe à relação argumentativa. Se houver função sintática, haverá uma relação argumental. Como num dicionário de regência (por exemplo, os dicionários de regência nominal e verbal de Celso Luft), em que a uma entrada são atribuídos seus complementos, no dicionário eletrônico, a todas as entradas seriam especificados os possíveis complementos. Se tudo o que é sintático se deve a relação argumentativa, assim, toda a sintaxe de uma frase estaria descrita pelas informações contidas nas entradas lexicais do dicionário eletrônico correspondentes as palavras da sentença.

Portanto, toda e qualquer formação sintática deve ser prevista na estrutura argumental de alguma entrada lexical. Dessa maneira, ao se desenvolver o dicionário eletrônico que acompanhará o programa de busca orientada a idéia, também se estará desenvolvendo a gramática.

Equiparar função sintática à relação argumentativa é simples com relação a algumas funções sintáticas. Um sujeito e um objeto podem ser facilmente entendidos como os argumentos de um verbo. Mas temos funções sintáticas tradicionais, cuja equiparação à relação argumentativa não é tão imediata assim.

Por exemplo, um artigo tradicionalmente recebe a função sintática de adjunto adnominal com relação ao substantivo ao qual se refere. No presente modelo, no entanto, não vamos considerar que as palavras gramaticais exerçam uma função sintática. Porém, devemos justificar a presença dessas palavras na estrutura sintática de uma sentença. As palavras gramaticais também serão parte de uma relação argumentativa.

O uso das palavras gramaticais, como os artigos e pronomes, é um fenômeno exclusivo da língua – não observamos um correspondente na realidade concreta. Devemos notar que a língua constitui uma realidade particular e como tal obedece a uma lógica intrínseca, apesar de receber influência de outras realidades, como visto[17].

Toda sentença de uma língua é formada a partir das regras gramaticais. Como sabemos, toda palavra que figura numa sentença, assim, tem sua existência justificada por alguma regra gramatical. Em nosso modelo não é diferente. Voltando às palavras gramaticais, devemos justificar o emprego delas numa sentença através da sintaxe. No entanto, sintaxe, aqui, é relação argumentativa. Então uma palavra gramatical deve participar de uma relação argumentativa.

[17] Por exemplo, a ordem das palavras é outro fenômeno exclusivo da língua, isto acontece, porque o substrato material no qual a língua se realiza, a fala, exige que as palavras sejam pronunciadas uma em seqüência de outra, enquanto na realidade os acontecimentos podem ocorrer simultaneamente. Por exemplo, precisamos pronunciar “João” antes de “matou” para expressar que foi João o assassino, e “José” depois de “matou” para expressar que foi José a vítima, no entanto, na realidade, não existe uma ordem entre os elementos dessa ação. João mata e José morre simultaneamente.

É fácil perceber que as preposições podem mediar uma relação argumentativa. Nas sentenças “adoro maçã” e “gosto de maçã”, o que difere no padrão argumental de seus verbos é que um dos argumentos do verbo “gostar” é introduzido por uma preposição. Para nós, é importante saber que o argumento pedido pelo verbo é o substantivo. Isso porque, como foi dito, a necessidade de um verbo de possuir um argumento nasce da mesma necessidade que uma ação tem de se realizar sobre a forma de entidades que a pratiquem. Como não há um correspondente concreto para uma preposição na realidade, então a preposição não é exatamente o argumento do verbo.

No entanto, na realidade lingüística existem preposições. E elas se justificam pela relação argumentativa de uma palavra. Afinal são apenas alguns verbos, como “gostar”, que pedem preposições. Além disso, essas palavras também selecionam as preposições que podem figurar entre elas e seus argumentos. Com o verbo “gostar”, por exemplo, somente pode figurar a preposição “de”, e não “para”, “em” ou “sobre”.

Fica claro que uma preposição participa de uma relação argumental. Mas por que está se tendo tamanho cuidado em enfatizar que, mesmo participando de uma relação argumental, uma preposição não será exatamente o argumento de um verbo, sendo visto apenas como uma intermediação entre o verbo e o substantivo que o complementa?

Essa é uma exigência prática do modelo, uma vez que ele se pretende servir a representar os significados de um texto. Como vimos, uma palavra que exerce função sintática em relação a outra implicará que as idéias que compõem seu significado vão especificar o significado da outra palavra. Sendo assim, toda função sintática deve estabelecer uma relação de especificidade entre as idéias das palavras participantes.

Ora, mesmo que mediados por uma preposição é entre os significados do verbo e do substantivo que se dá a relação de especificação. Então deve ser estabelecida entre verbo e substantivo a função sintática. Esse é um ponto.

Outro motivo é que representamos as relações de especificação entre idéias por um grafo. Como se sabe, um arco de um grafo liga dois e somente dois elementos. Então temos que definir somente o substantivo como complemento do verbo e não a

preposição e o substantivo como complementos, porque assim teríamos três elementos participando de um arco de um grafo.

Assim como as preposições, as demais palavras gramaticais numa sentença devem ser justificadas a partir de uma relação argumental.

A presença dos pronomes também pode ser explicada dessa forma. Veja o caso do texto “Falei com José. Ele virá hoje”.

O pronome “ele” não é visto neste modelo como o complemento do verbo “vir”. Este verbo pede como seu complemento um substantivo. Para tanto, uma possível complementação a esse verbo pode ser feita de maneira indireta através de um pronome. O pronome “ele”, assim, intermedeia a relação argumental estabelecida entre o verbo “vir” e o substantivo “José”.

Essa mecânica do modelo não é apenas um exercício formal de reduzir a sintaxe à regência. No caso do pronome, fica claro que essa mecânica é uma habilidade indispensável ao modelo, justamente por ele visar a representação das paráfrases. O exemplo mostra que o verdadeiro agente do evento “vir” em “Ele virá hoje” não é o pronome “ele”, mas é “José”. Essa sentença pode ser parafraseada pela sentença “José virá hoje”. É fundamental, portanto, que a ligação seja feita entre o verbo e o substantivo, sendo que o pronome anafórico intermedeia essa ligação indireta[18]. Somente assim podemos encontrar o significado correto dessa construção, isso é, sua paráfrase correta.

O fato de estarmos trabalhando com o grafo nos obriga, como dito anteriormente, a definir a função sintática como sendo uma relação entre duas palavras. Desse modo, o modelo que estamos construindo encontra um problema com um recurso muito usado nas teorizações gramaticais: o sintagma.

[18] Provavelmente essa intermediação entre o verbo e o substantivo não será dada apenas pelo pronome. O uso da pontuação é fundamental para nós leitores atribuirmos ao substantivo “José” o papel de sujeito do verbo “virá”. Em trabalhos futuros, as pontuações (ponto final, vírgula, parêntesis etc.) também poderão ser tratadas como um elemento de intermediação numa relação argumental.

Ao invés de considerar função sintática como algo que ocorre entre duas palavras, as teorias que fazem uso do sintagma consideram que uma função sintática ocorre entre uma palavra e um bloco de palavras, o sintagma[19].

Isso não significa que a parte gramatical desse modelo tenha que obrigatoriamente banir o recurso ao sintagma. Temos duas opções: ou começamos a teorizar uma nova sintaxe otimizada para nossos fins e que, portanto, desde os princípios de suas teorizações atendam às exigências da representação dos significados, isso é, que limite a função sintática à relação argumental, que considere a relação argumental como sendo algo relacionado a duas palavras e que trate as palavras gramaticais como apenas um intermédio das relações argumentais; ou, então, podemos aproveitar os avanços já conseguidos com as teorizações que incluem o sintagma, adaptando-as às exigências do modelo de representação dos significados.

Por exemplo, para podermos atribuir a relação de especificação entre as estruturas de conceitos das palavras que se relacionem sintaticamente, no modelo que admitisse o sintagma, teríamos de fazê-lo entre os núcleos dos sintagmas.

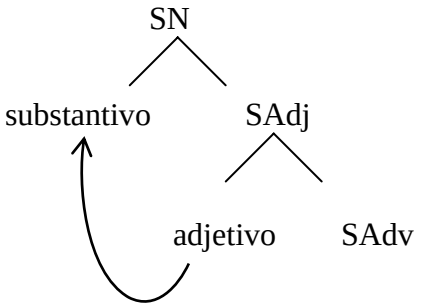


Figura 26 – núcleos do sintagma adjetivo e nominal

Outro cuidado a ser tomado é com as palavras gramaticais, que não participariam diretamente da relação de especificação, e, portanto, o processo de relacionar sintaticamente os núcleos dos sintagmas não poderia ser levado em conta quando o núcleo de um sintagma for uma palavra gramatical. É o caso do sintagma preposicional.

[19] Isso ocorre inclusive na teoria X-barra de Chomsky em *Princípios e Parâmetros*, pois, a pesar de o teórico sugerir uma árvore sempre binária que, portanto, liga dois elementos, esses elementos nunca são duas palavras.

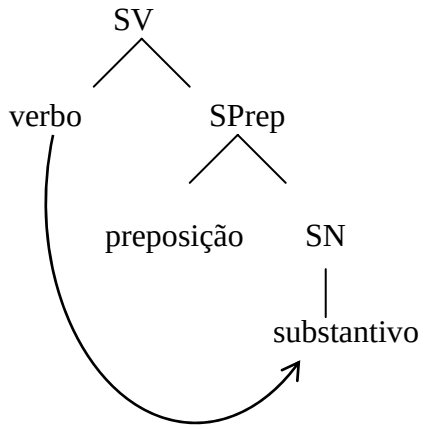


Figura 27 – ignorando o núcleo do sintagma preposicional

Além disso, existe uma ordem hierárquica entre os sintagmas que teria de ser subvertida durante a adaptação ao modelo de representação semântica. Por exemplo, aquele sintagma nominal que funciona como o sujeito está em posição hierárquica igual à do sintagma verbal – pois ambos são elementos constitutivos do sintagma oracional.

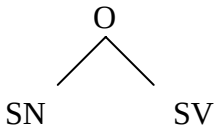


Figura 28 – hierarquia entre sintagma nominal do sujeito e sintagma verbal

No entanto, o sintagma nominal correspondente ao objeto direto está em posição hierárquica inferior ao sintagma verbal – isso porque o sintagma nominal do objeto direto é parte constitutiva do sintagma verbal. Desta maneira, o sintagma nominal do sujeito e o sintagma nominal do objeto direto estão em posições hierárquicas diferentes em relação ao verbo.

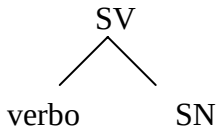


Figura 29 – hierarquia entre sintagma nominal do objeto e sintagma verbal

Mas do ponto de vista da relação argumental, o sujeito e o objeto, entre si, possuem a mesma posição hierárquica – ambos são argumentos do verbo.

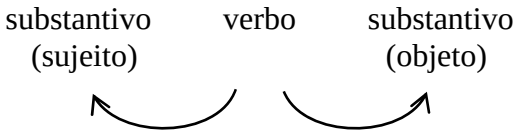


Figura 30 – hierarquia entre os argumentos do verbo

Além disso, a posição hierárquica entre a palavra regente e seu argumento nem sempre é a mesma que a posição hierárquica estabelecida entre os núcleos dos sintagmas correspondentes à palavra regente e ao argumento. Por exemplo, o verbo está numa posição hierárquica superior ao substantivo que é núcleo do sintagma nominal correspondente ao objeto direto.

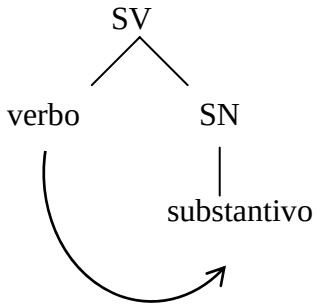


Figura 31 – hierarquia entre verbo e seu argumento

Mas o adjetivo, núcleo do sintagma adjetivo, está em posição hierárquica inferior ao substantivo, núcleo do sintagma nominal do qual o sintagma adjetivo também é parte constitutiva. Então a posição hierárquica não pode ser levada em consideração para atribuir o sentido da relação argumentativa.

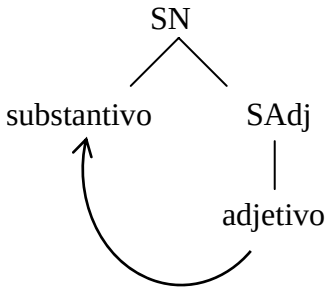


Figura 32 – hierarquia entre o adjetivo e seu argumento

Por considerar que o esforço de adaptar as teorias que admitem o sintagma é maior do que teorizar do zero uma gramática simplificada que atendesse exatamente às exigências da representação dos significados, será dada a preferência nessa pesquisa pela segunda opção.

Mas será que excluir o sintagma comprometerá de alguma forma a gramática do modelo que pretendemos construir?

Não há necessidade da existência de sintagmas porque as necessidades que o sintagma vem a suprir nos modelos que o adotam podem ser facilmente resolvidas de outra maneira nesse modelo.

Os sintagmas existem num modelo basicamente para tratar três questões: adjuntos, substituições e recursividade.

Vejamos um exemplo em que a figura do adjunto aparece: “O gato gordo comeu muita coxa de galinha”. Alguns modelos diriam que o verbo “comer” não pede um substantivo como um dos seus complementos, mas um sintagma substantivo. Esse sintagma substantivo seria reescrito por um sintagma composto pelo determinante, seguido por um substantivo (“gato”), seguido por um sintagma adjetival. O determinante seria reescrito pelo artigo (“o”) e o sintagma adjetival, pelo adjetivo (“gordo”).

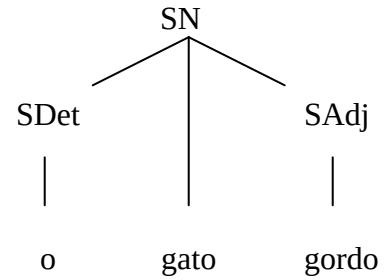


Figura 33 – estrutura do sintagma substantivo

O que nos importa aqui é que um dos complementos do verbo “comer”, pelo modelo que se serve do sintagma, é “o gato gordo”.

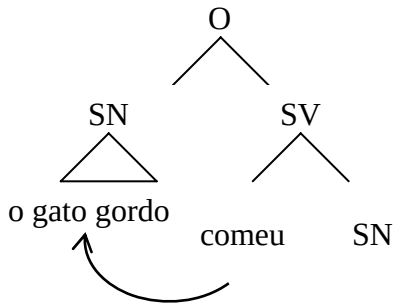


Figura 34 – sintagma nominal como argumento do verbo

Quando o sintagma é excluído, no entanto, apenas o substantivo “gato” pode ser um dos complementos do verbo “comer”, sendo que essa ligação é intermediada pelo artigo “o”[20]. Todas as seleções feitas pelo verbo a seu argumento – seleção da classe de palavra e das propriedades semânticas do argumento – somente são feitas com relação ao substantivo. Sabemos que a ação de comer somente pode ser executada por uma entidade. As entidades são representadas na língua pelo nome. Portanto o verbo “comer”, ou mais propriamente o evento “comer”, exige que seu complemento seja um nome. E mais do que isso, a ação de comer somente poderá ser executada por uma entidade com certas características específicas. Essa entidade

[20] É sempre importante lembrar que uma palavra gramatical, como é o caso do artigo “o”, funciona apenas como elemento intermediário de uma relação argumentativa.

deverá ser um ser animado[21]. Essa mesma seleção é transposta para a relação argumental estabelecida entre o evento "comer" e o nome "gato"[22].

Portanto, a relação argumental somente se preocupa com a palavra "gato". Nessa relação argumental é irrelevante saber sobre o adjetivo "gordo". A existência desse adjetivo na sentença se faz por outra relação argumental, a relação estabelecida entre o adjetivo "gordo" e o substantivo pedido por ele, "gato".

Assim, um adjunto como "gato", não precisará da figura do sintagma para ser tratado porque não é pela relação argumental existente entre o verbo e o sintagma nominal inserido no sintagma verbal que se justificará a presença do adjunto na sentença. Um adjunto possui sua própria relação argumental[23]. Já podemos dispensar o sintagma para a questão do adjunto.

Outra questão tratada pelo sintagma é a substituição. A substituição é o que ocorre, por exemplo, quando um sintagma nominal pode ser reescrito como um pronome anafórico. Veja a construção:

A coxa de galinha estava na mesa. O gato comeu *isso*.

Segundo a teoria que admite o sintagma, um sintagma pode ser reescrito de diversas formas. No caso do sintagma nominal, a reescrita tanto poderia ser feita da forma como explicado no exemplo anterior, quanto poderia ser feita por um pronome, como acontece com o pronome "isso" desse exemplo. É essa construção "alternativa" que estou chamando de substituição.

[21] Se imaginarmos outro significado para a palavra "comer", não estamos nos referindo à mesma ação. Pode-se afirmar, sem receios, que, nesse sentido, a ação comer é executada sempre por um ser animado.

[22] Veremos no próximo tópico como se dá essa seleção semântica na estrutura de conceitos do argumento.

[23] Os demais adjuntos (se não forem palavras gramaticais, como os artigos), inclusive as orações adjetivas, devem ser tratadas dessa forma. Mas como o relacionamento sintático, aqui, não pode se dar entre uma palavra e um bloco de palavras (a oração) esse relacionamento será um tanto diferente. O verbo da oração adjetiva pede como argumento o substantivo a qual incide na oração dita principal, sendo que o pronome relativo é um intermédio dessa relação argumentativa.

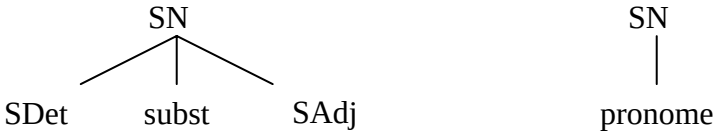


Figura 35 – substituição no sintagma nominal

O modelo sem sintagma tratará o fenômeno da substituição de outra maneira. O pronome "isso" se refere a um substantivo. Nesse modelo, a referência que um pronome faz a um substantivo será tratada como uma ligação indireta, isto é, o pronome intermediará a ligação entre a palavra que pede o substantivo como seu argumento e o próprio substantivo. Nesse caso, o pronome "isso" somente poderá se ligar a "coxa", pois o verbo comer faz uma seleção semântica de seu complemento, impossibilitando a ligação com "mesa" [24].

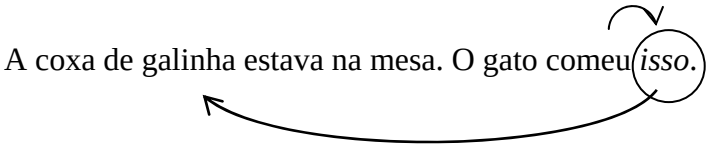


Figura 36

Dessa forma, não precisamos recorrer ao sintagma para o tratamento das substituições.

A última questão tratada pelo sintagma é a recursão. A recursão de um sintagma acontece quando a reescrita desse sintagma for feita por, dentre outros elementos, o próprio sintagma.

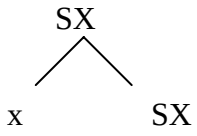


Figura 37 – estrutura de um sintagma com recursão

[24] Outras regras impedem que "galinha" (ordenação) e "gato" (tipo de pronome) sirvam como complementos.

Nos nossos exemplos, “muita coxa de galinha” representa um sintagma nominal em que aparece uma recursão. Nesse caso, tem-se o que se chama de sintagma preposicionado. Um sintagma preposicionado nada mais é que uma preposição seguida de um sintagma nominal. Dessa forma, o sintagma nominal “muita coxa de galinha” contém outro sintagma nominal, “galinha”.

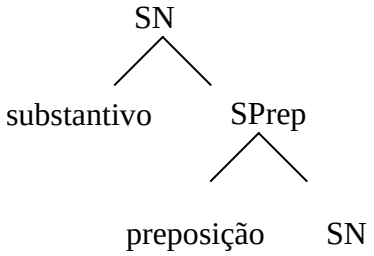


Figura 38 – recursão no sintagma nominal

Esse segundo sintagma nominal poderia ser reescrito de diferentes formas, com uso de artigos, adjetivos, outros sintagmas nominais etc.

O recurso ao sintagma, portanto, facilitaria a explicação das reescritas, pois não seria necessário explicar todas as palavras que poderiam aparecer num sintagma (um trabalho impossível, tratando-se das recursões), mas bastaria que um sintagma encapsulasse todas as alternativas de construção.

O presente modelo simplificará ainda mais esse processo. Um sintagma preposicionado nada mais é que um argumento de um substantivo. Esse tipo de argumento é sempre pertencente à classe dos substantivos e sempre mediado por uma preposição. Um argumento é tratado da maneira como o viemos tratando até agora, então podemos continuar a fazer isso sem recorrermos ao uso dos sintagmas. Não precisamos prever uma estrutura como o sintagma nominal que em alguns casos seja composto por uma recursão (o sintagma preposicionado). Somente no caso do substantivo analisado necessitar de um complemento (e isso se sabe ao consultar o dicionário) é que essa informação será levada em conta. Simplificamos assim nosso trabalho.

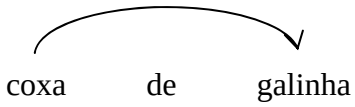


Figura 39 – sintagma preposicionado como argumento do substantivo

Outras construções recursivas também podem ser tratadas através das relações argumentais. Por exemplo, a oração subordinada substantiva é uma oração que compõe a oração principal, segundo o modelo dos sintagmas. Isso quer dizer que na estrutura do sintagma oracional pode aparecer um outro sintagma oracional. Quando isso ocorre, há uma recursão.

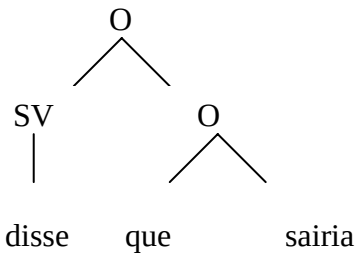


Figura 40 – recursão no sintagma verbal

Segundo o modelo proposto, no entanto, o verbo da oração considerada principal pede o verbo da oração considerada subordinada substantiva como seu complemento, sendo que essa ligação é intermediada pela conjunção.

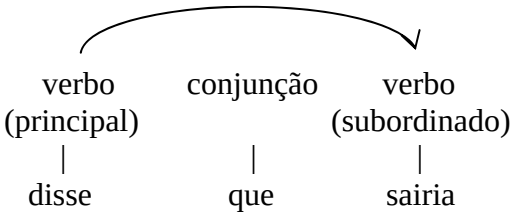


Figura 41

Outros tipos de orações subordinadas substantivas também se valem desse mesmo processo.

Assim, os casos tratados como recursão no modelo que considera o sintagma também podem ser tratados sem esse recurso.

O mais interessante de não recorrer ao sintagma é que estamos usando justamente as próprias exigências[25] do modelo de representação dos significados para justificar os casos resolvidos pelo sintagma. Assim, a teoria fica mais enxuta e coerente com suas propostas básicas.

Finalmente podemos passar para a questão levantada pelo título desse tópico: como formalizar os argumentos do dicionário eletrônico?

Os significados puderam ser representados por grafos, porque a relação de especificação entre significados foi comparada com a relação argumentativa estabelecida entre palavras. Assim, as estruturas de conceitos das palavras que estabeleciam relações argumentativas poderiam ser unidas numa única estrutura de conceitos comum às duas palavras, por reproduzir a ligação sintática existente entre duas palavras.

Então podemos representar as relações argumentativas da mesma forma como representei as relações conceituais: através de grafos. Desse modo teremos uma mesma forma de tratamento para a sintaxe e para a semântica: o grafo.

Ao tratarmos semântica e sintaxe com um mesmo recurso, isso não estaria apenas mantendo a teoria mais coesa, mas poderíamos aproveitar a agilidade computacional de se trabalhar com grafos tanto para tratar a sintaxe quanto para tratar a semântica.

No dicionário eletrônico, a informação a respeito das relações argumentais será expressa, portanto, como um grafo. Esse grafo será chamado no presente modelo de estrutura de argumentos[26]. A representação formal da estrutura de argumentos de uma entrada lexical deverá obedecer ao seguinte esquema:

[25] As exigências são: i) equipar toda função sintática a uma relação argumental, ii) a relação argumental só pode agir sobre duas palavras, e iii) as palavras gramaticais só podem ser vistas como intermédio de relações argumentais.

[26] O termo “estrutura de argumentos” foi retirado de Pustejovsky (1995), apesar de não possuir aqui o mesmo conceito que em sua obra.

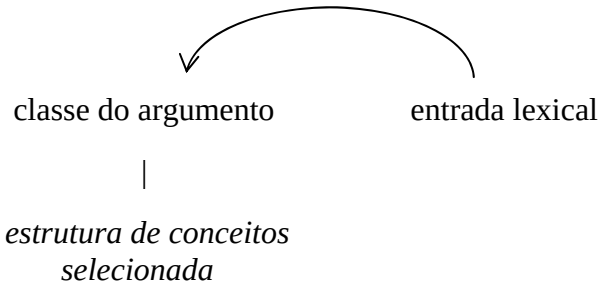


Figura 42 – representação da estrutura de argumentos

Existem três partes nesse representação, sendo que a parte central é composta pela própria entrada lexical. Esta se liga por um arco à classe de palavra selecionada pela entrada para o seu argumento. O argumento se liga com uma seta reta não direcionada à outro grafo que compõe a estrutura de conceitos selecionada pela entrada para o seu argumento. Vejamos como isso se dá com exemplos:

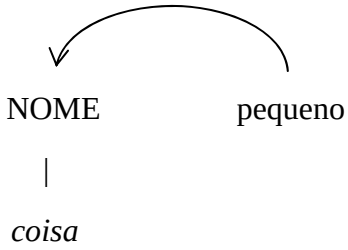


Figura 43 – exemplo de estrutura de argumento

Como ainda não nos detivemos nas estruturas de conceitos, estão sendo usados, por isso, conceitos hipotéticos representados por letras. Lembre-se ainda que um grafo pode ser composto de um único elemento, como é o caso da estrutura de conceitos do exemplo, composto apenas pelo conceito “coisa”. Vejamos outro exemplo:

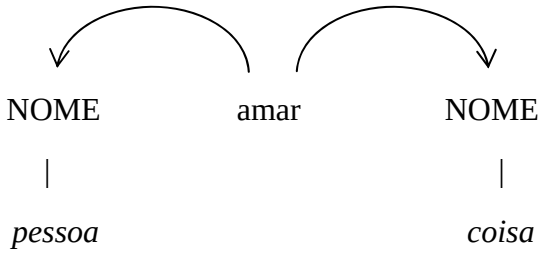


Figura 44 – seleção semântica

A estrutura de conceitos deverá apresentar todos os argumentos pedidos pela entrada lexical em um único grafo.

Na representação gráfica com a ligação intermediada por outras palavras, um arco não ordenado une a palavra regente à palavra intermediária e um arco ordenado (uma seta) une a palavra intermediária, finalmente, ao argumento. A estrutura básica é a seguinte:

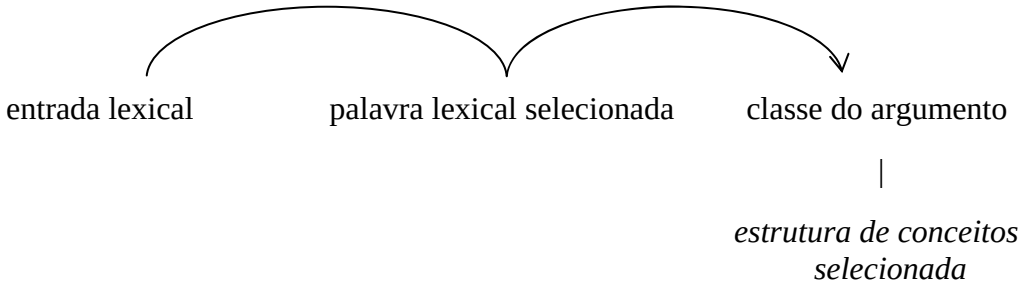


Figura 45 – representação da seleção semântica na estrutura de argumentos

Eis um exemplo disso:

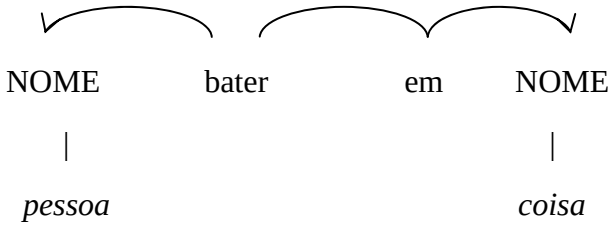


Figura 46 – exemplo de seleção semântica

Como pode ser notado, se a entrada selecionar uma palavra gramatical específica, como o verbo seleciona a preposição “em” e não outra preposição qualquer, então essa palavra deve ser especificada na estrutura de argumentos; se selecionar somente a classe de palavra que intermediará a ligação, então assim deverá ser representada. No caso das preposições, essa classe não gera estrutura conceitual, por isso também não será selecionada uma estrutura conceitual. Mas algumas palavras gramaticais, como veremos mais adiante, produzem estruturas conceituais e, portanto, também delas são selecionadas estruturas de conceito.

Tínhamos optado pelo grafo para as estruturas de conceito justamente por ele permitir a flexibilidade de não nos importarmos com a ordem dos conceitos. Na teoria dos grafos, um grafo se diferencia de outro apenas pelos nós e pelas relações entre os nós. Portanto esses grafos são congruentes:

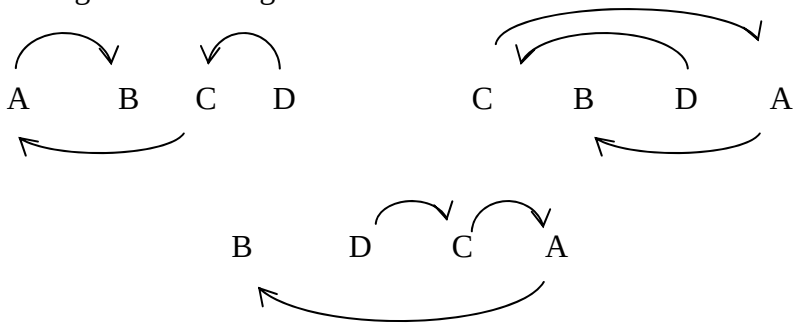


Figura 47 – grafos congruentes

Isso é extremamente útil quando esses grafos estiverem representando a estrutura de conceitos, tão útil que vamos continuar optando pelos grafos para formalizar as estruturas de conceitos e conseqüentemente também será formalizada através de grafos a estrutura de argumentos. Entretanto, os grafos nas estruturas argumentais ligam palavras, e sabemos que, se esses mesmos nós dos exemplos acima estiverem representando palavras, essa liberdade não corresponde à realidade.

Duas sentenças que apresentem as mesmas palavras, mas as ordenem de forma diferente podem ter seus significados diferentes. Era o caso de “João matou José” e “José matou João”. Além disso, sabemos que uma palavra não pode ocupar qualquer lugar na sentença. Por isso, o modelo gráfico de representação da estrutura de argumentos deverá possuir técnicas para controlar a flexibilidade própria dos grafos.

As palavras devem obedecer a uma disposição na sentença. Tem-se basicamente três problemas referentes à disposição das palavras numa sentença: i) deve-se saber se o argumento antecede ou sucede a palavra que o pede; ii) quando uma palavra pede mais de um complemento, é necessário saber a ordem deles; iii) quando uma palavra serve de argument Uma ordenação topológica possível:
de todas essas palavras.

Existe um recurso nos estudos dos grafos que se chama ordenação topológica. A ordenação topológica visa unificar grafos que compartilhem alguns elementos num único grafo em que cada elemento apareça ordenado numa seqüência linear. O exemplo clássico é um grafo que representa a ordem de se vestir as peças de roupa, em que as setas apontam a peça que deve vir depois de outra:

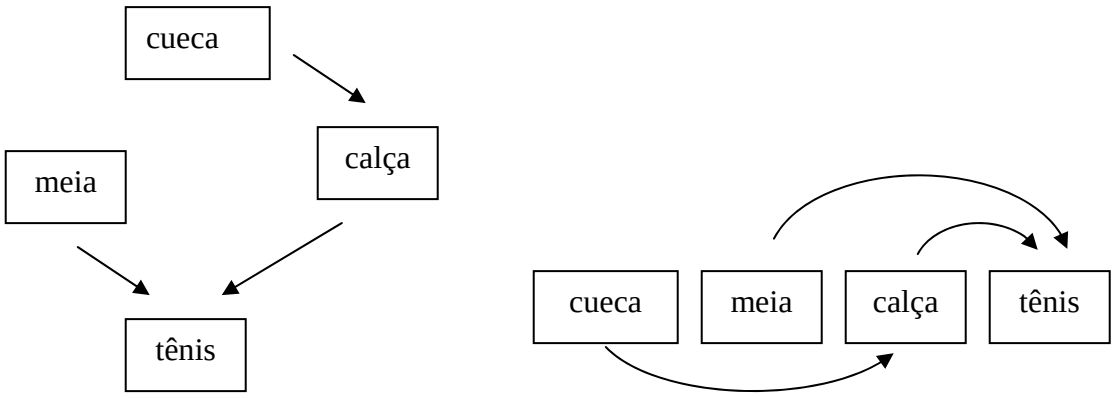


Figura 48 – ordenação topológica

Um arco de um grafo pode representar qualquer relação que possa se estabelecer entre os elementos, contanto que seja a mesma relação. A particularidade determinante na ordenação topológica é que as setas sempre representam o mesmo tipo de relação: seqüências. Podem ser seqüências de diferentes naturezas, temporais, espaciais etc., mas são sempre seqüências. E por isso, em todo grafo resultante de uma ordenação topológica, as setas sempre apontam para um mesmo sentido. Veja-se que, no nosso exemplo, as setas do grafo ordenado apontam para a direita.

Mas isso não ocorre com os arcos dos grafos em noso modelo. Nos grafos que representam os significados, as setas fazem as vezes da relação de especificação estabelecida entre os conceitos e, nos grafos que representam as regências, as setas

fazem as vezes da relação argumental. Como essas relações representadas pelos arcos do grafo não são seqüências, as setas em ambas as estruturas podem apontar para a direita ou para a esquerda.

No entanto, as estruturas argumentativas participam de um processo de que a ordenação topológica participa: a unificação de vários grafos num único grafo em seqüência linear de seus elementos. Portanto parece natural e desejável recorrermos a essa técnica amplamente estudada e já desenvolvida.

Se quisermos fazer uma ordenação topológica, devemos recorrer a um outro esquema de setas que aja concomitantemente com o esquema que represente as relações argumentais. Esse novo esquema representará a sucessão dos elementos.

Usaremos uma seta reta e com a ponta cheia (como nos exemplos abaixo) sempre apontando para a direita, para representar a sucessão de elementos. Então, visando solucionar o problema (i), usaremos a seta da ordenação topológica ligando argumento a palavra que o pede, se o argumento o anteceder, ou usaremos a seta ligando a palavra que pede o argumento a ele, se o argumento suceder a palavra que o pede. Dessa forma:

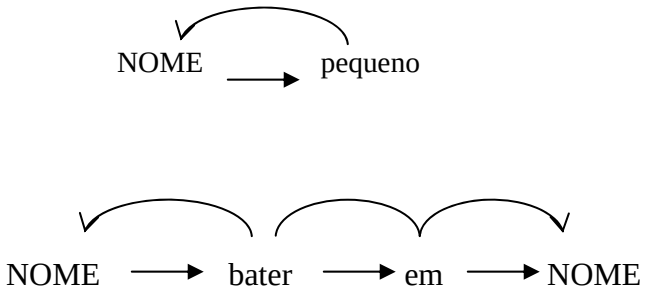


Figura 49 – ordenação topológica na estrutura de argumentos

3. 2. 2
A estrutura de conceitos das palavras

Inicialmente, serão lembrados alguns pontos já discutidos anteriormente. A incumbência desse modelo de representação é permitir que o computador determine qual a estrutura de conceitos que forma o significado de um texto para que, ao invés de um sistema de busca procurar pelas palavras escritas no texto, procure pelas idéias que delas se formam. O significado do texto será representado por uma enorme

estrutura de conceitos que é composta pelas estruturas de conceitos de cada palavra presente no texto, unidas pelas relações sintáticas estabelecidas entre estas palavras.

A estrutura de conceitos de cada palavra é gerada a partir da definição de qual sentido a palavra está exercendo na sentença que ela ocupa. A definição de sentidos se faz por meio da análise sintática.

No léxico eletrônico que acompanha o programa que irá gerar as estruturas de conceitos dos textos, cada significado diferente de uma palavra deverá corresponder a uma entrada distinta. Cada entrada deverá informar a qual classe gramatical a palavra corresponde, quais argumentos ela pede e a qual estrutura de conceitos corresponde seu significado. Com essa informação será possível fazer a análise sintática da sentença, para então definir o sentido de cada palavra e, finalmente, gerar a estrutura conceitual.

A estrutura de conceitos de uma palavra é um grafo formado por conceitos (os nós) e as relações de especificação entre esses conceitos (os arcos). As relações de especificação estabelecidas entre conceitos guardam semelhança com as relações sintáticas estabelecidas entre as palavras, também representadas por grafos.

Os conceitos da estrutura de conceitos de uma palavra são definidos para dar conta do significado dessa palavra. Veja-se, no entanto, que “significado” não é exatamente a “compreensão” que temos dessa palavra. Como já dissemos, definir o significado de uma palavra é simplesmente definir quais paráfrases podemos fazer a partir dela.

Paráfrases têm o mesmo significado e, portanto, a mesma estrutura de conceitos.

Dessa maneira, os conceitos são definidos tendo em vista as outras palavras ou expressões por que podem ser substituídas.

A estrutura de conceitos, que representa um significado de uma palavra, será levantada no momento da feitura do dicionário eletrônico que acompanhará o sistema de busca orientado a idéia.

E como definiremos quais conceitos e quais relações entre esses conceitos comporão a estrutura conceitual que virá a representar o significado de uma dada palavra? O procedimento para tanto seria levantar todas as construções em que

apareça a tal palavra no sentido pretendido e cotejá-las com todas as possíveis formas que as parafraseiem[27]. Dessa maneira definiremos o número de conceitos e quais são eles, além das relações estabelecidas entre esses conceitos.

No entanto, cada paráfrase também pode ser parafraseada. E cada paráfrase da paráfrase também pode ser parafraseada. Por isso, a definição do significado de uma única palavra dependerá muitas vezes de um enorme número de outras palavras da realidade lingüística em que estão inseridas. Portanto, para conseguirmos realizar o procedimento total de cotejar uma palavra com todas as paráfrases que ela pode fazer numa realidade lingüística, necessitaríamos de um léxico e uma gramática com extensões próximas a de usos reais da língua.

Para tanto, despenderíamos um trabalho hercúleo (devendo ser automatizado, inclusive), fugindo assim da extensão e propósitos de uma dissertação. Dessa maneira, as estruturas de conceitos apresentadas no presente trabalho são apenas hipotéticas, porque, da forma como são representadas aqui, não poderiam ser usadas por um sistema de busca que incida sobre textos produzidos numa realidade lingüística verdadeira.

Ao apresentarmos aqui nossa proposta de um modelo que represente a semântica voltado para a busca orientada a idéia, nos limitaremos apenas em demonstrar a metodologia que futuras análises deverão seguir a fim de construir o dicionário eletrônico de um sistema de buscas por idéias. Assim sendo, devemos nos ater a certas técnicas de decomposição do significado em estruturas conceituais.

[27] Evidentemente, é impossível determinar todas as paráfrases se considerarmos a criatividade lingüística. Um falante de uma língua sempre pode imaginar uma nova construção para expressar uma determinada idéia ou um novo sentido para uma palavra. Esse é um problema também enfrentado sempre que se desenvolve um dicionário ou uma gramática, por isso sempre há a necessidade de atualizar ambos. No entanto, como acontece sempre que se define um dicionário, estamos limitando o sentindo que cada palavra e, portanto, cada expressão exercem. Se nunca admitirmos que podemos levantar todos os sentidos de uma palavra ou levantar todas as palavras que podem exercer uma idéia, supondo que a criatividade lingüística impedirá tal feito, é melhor nem começar a criar um buscador orientado a idéia, como também é melhor nem começar a escrever um dicionário, ou, ainda, é melhor nem começar a propor uma teoria semântica. Pelo bem do desenvolvimento tecnológico e intelectual, podemos afirmar, sem receios, que, dentro do universo de sentidos delimitados pelo estágio atual da língua e limitado pela competência lingüística de quem se empenhar na representação dos significados, é possível levantar todas as paráfrases possíveis para expressar uma idéia. O buscador, assim como um dicionário, terá suas falhas, mas é inegável que haverá utilidade.

Por exemplo, os conceitos que compõem um nome devem levar em conta dois aspectos semânticos, visando as paráfrases: a sinonímia e a relação entre hiperônimos e hipônimos.

Vamos imaginar que um programa que defina as estruturas de conceitos se utilize de um léxico cujo universo de palavras seja composto apenas pelos nomes: fruta, maçã, banana, tangerina, mexerica.

Os nomes “maçã”, “banana”, “tangerina” e “mexerica” são hipônimos de “fruta”. Os hipônimos são hiperônimos especificados. Por exemplo, uma banana é uma fruta com algumas características que a especificam e, portanto, a diferenciam de outras frutas como as maçãs. Se existe relação de especificação, isso deve ser representado na estrutura de conceitos. Então, toda a estrutura de conceitos que venha a compor “fruta” deve estar contida na estrutura de conceitos de seus hipônimos como subgrafos.

As estruturas de conceitos apresentadas na figura a seguir têm conceitos hipotéticos:

Figura 50 – hiperônimo e hipônimo

A definição da estrutura de conceitos que representa o significado de um nome precisa trazer toda a taxionomia que envolve esse nome. São dois os motivos para isso. O primeiro diz respeito à seleção semântica que os eventos e qualificadores fazem dos nomes que lhes servem de argumento; veremos isso no próximo tópico.

O outro motivo diz respeito às paráfrases realmente. Observe-se essas duas construções: “O menino comeu uma banana” e “O menino encontrou uma banana na pia. Ele comeu a fruta”.

Um hiperônimo tem a capacidade de fazer referência a um hipônimo, assim como o pronome faz referência a um nome. No contexto em que está inserida, “Ele comeu a fruta” é absolutamente uma paráfrase de “O menino comeu uma banana”.

Para que o programa atribua “banana” corretamente como argumento de “comeu”, é necessário representar a relação hiperônimo/hipônimo na estrutura de conceitos.

A partir daí, podemos descrever as estruturas de conceitos dos nomes. Se o significado de uma palavra é levantado a partir das possíveis paráfrases que se podem fazer dela, então as estruturas de conceitos que compõem o significado das palavras são definidas a partir das possíveis paráfrases que podemos fazer com as palavras e expressões existentes.

Portanto é muito importante, durante a definição das estruturas de conceitos das entradas lexicais do dicionário eletrônico, saber, primeiro, quais as reais paráfrases que se formam no corpus para o qual esse dicionário será desenvolvido, e, segundo, quais são as palavras existentes no léxico para representar essas palavras[28]. O lingüista deve fugir a detalhamentos inúteis nas estruturas de conceitos. Não é necessário definir um conceito se esse não servir às construções corretas de paráfrases do uso estudado. Isso quer dizer que se no léxico em questão não existirem as palavras “mimosa” e “bergamota”, a palavra “tangerina” só poderá ser parafraseada por “mexerica”. Não podemos parafrasear “banana” por “fruto da bananeira, oblongo e de polpa carnosa, sem sementes, desenvolvido através de cultura, mais ou menos recurvado, com casca verde e, quando maduro, amarela, parda ou avermelhada, com polpa branco-amarelada ou amarela, pastosa, doce, aromática, espécie rica em amido e potássio”[29] se no corpus não houver momento em que banana apareça parafraseado assim.

Se, em nosso léxico, os nomes fossem apenas os citados (fruta, maçã, banana, tangerina, mexerica), então as estruturas de conceitos seriam as seguintes:

[28] Conhecer todas as palavras do corpus é algo simples de fazer, basta usar um programa elementar para listar todas as palavras diferentes que aparecem num corpus. Já prever todas as paráfrases que expressem uma idéia presentes no uso estudado é algo que um lingüista com intuição razoavelmente apurada e sem receios de tomar decisões pode fazer tranqüilamente.

[29] Definição de “banana” dada pelo dicionário *Houaiss*.

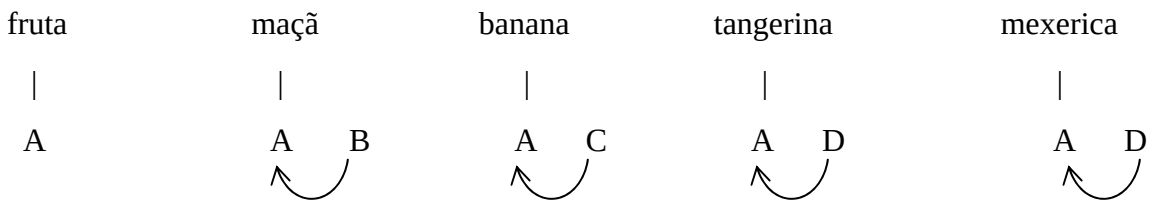


Figura 51 – exemplo de atribuição de conceitos

No exemplo exposto acima, já estão presentes alguns procedimentos importantes. Primeiro, definimos apenas um conceito para fruta, representado pela letra A. Mas há algumas páginas atrás, tínhamos dito que um nome deveria trazer toda a taxionomia que a antecedesse. Por que não fizemos isso para a palavra fruta?

Ora, no nosso léxico demonstrativo não existem palavras que sejam hiperônimos de fruta. Se não existem no léxico, não precisam existir na estrutura conceitual, porque não existirão paráfrases para elas.

Os outros nomes, por serem hipônimos de “fruta”, possuem em suas estruturas toda a estrutura conceitual de “fruta” que, no caso, é apenas o conceito A. Elas se diferenciam entre si e também do conceito de “fruta” pelo único conceito acrescentado à estrutura de conceitos de “fruta” ligada por uma relação de especificação.

As palavras “mexerica” e “tangerina” são sinônimas e, portanto, a estrutura conceitual de ambas é idêntica.

Até agora em nossos exemplos os conceitos vinham sendo representados por letras do alfabeto grafadas em letras maiúsculas. Mas, num universo real de uso da língua, é muito provável que esse recurso seja muito ruim, já que, muito provavelmente, existirão mais conceitos do que as letras do alfabeto, além de esse ser um recurso bem pouco mnemônico.

Por isso a partir de agora tenderei a escolher formas mais mnemônicas para designar um conceito.

Em suma, o arranjo básico da estrutura de conceitos é o seguinte:

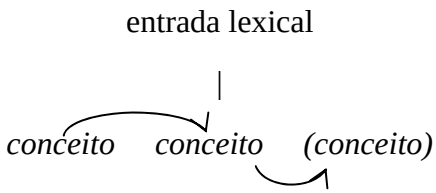


Figura 52 – representação da estrutura de conceitos

Note-se que um dos conceitos apareceu entre parêntesis. Lembre-se que na estrutura de argumentos de uma palavra era selecionado um conceito que deveria figurar na estrutura de argumentos de seu argumento (como foi exemplificado na figura 44).

Como no momento em que foi apresentada a estrutura de argumentos ainda não havíamos exposto a estrutura de conceitos, foram demonstrados exemplos em que uma palavra regente selecionava apenas um conceito de seu argumento. No entanto, é selecionada, na verdade, uma pequena estrutura de conceitos – que pode ser composta por um ou mais conceitos ligados entre si por relação de especificação. Isso quer dizer que, para que uma palavra possa servir de argumento a outra, ela deve possuir como um subgrafo de sua estrutura de conceitos o pequeno grafo pedido pela palavra regente. Assim se faz a seleção semântica dos argumentos neste modelo.

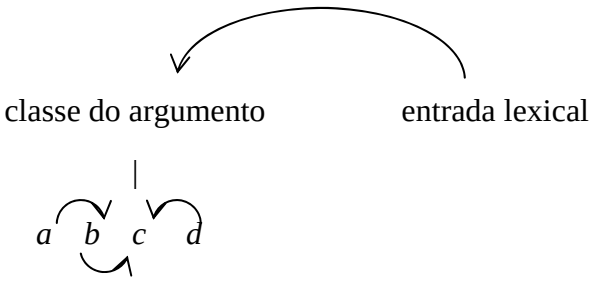


Figura 53 – estrutura de argumentos selecionando um grafo

Quando foi apresentado o esquema básico da estrutura de conceitos, um dos conceitos estava entre parêntesis. Aquele conceito que veio entre parêntesis é o conceito pedido por uma palavra regente.

É importante notar que, na estrutura de conceitos da palavra regente, o subgrafo que representa as idéias selecionadas para constar na estrutura de conceitos de seu

argumento também faz parte da estrutura de conceitos da palavra regente. Quer dizer o significado de uma palavra regente é composto também por parte do significado de seus argumentos[30]. No entanto, na estrutura de conceitos da palavra regente, esses conceitos deverão estar destacados entre parênteses, porque no momento de unificar as estruturas de conceitos da palavra regente e do argumento, as estruturas entre parêntesis não podem se repetir. Isso ficará mais claro quando exemplificarmos no capítulo 4.

3. 2. 3
O componente textual do modelo

Como foi dito desde o início da exposição deste modelo de representação, um texto será substituído por um grande grafo que representa as idéias expressas nele. Esse grafo nada mais é do que os pequenos grafos das estruturas de conceitos de cada palavra do texto unidos pelas relações sintáticas estabelecidas pelas palavras. No entanto, o leitor mais cuidadoso terá percebido que a sintaxe somente ligará as estruturas das palavras contidas numa mesma sentença. Por isso, o grafo total do texto poderá ser um grafo desconexo em que os vários subgrafos formados pelas sentenças não se conectam.

Um grafo desconexo é um grafo em que, se partindo de um de seus elementos, não se pode chegar a qualquer outro seguindo os arcos. Como é definido, o conceito de desconexão de um grafo pode parecer complicado, mas, notando-se exemplos ilustrativos de grafos conexos e desconexos é bem fácil entender:

[30] É curioso notar que, em alguns momentos, os dicionários se utilizam desse princípio nas definições de palavras que precisam de complemento, principalmente com verbos e adjetivos. Por exemplo, quando a definição de “recear” é dada por “ter medo de alguma coisa ou alguém” ou quando “receoso” se define por “aquele que tem medo de alguma coisa ou alguém”, está-se incluindo nas definições dos sentidos dessas palavras certas propriedades semânticas dos seus complementos através das palavras “alguém”, “alguma coisa”, “aquele”.

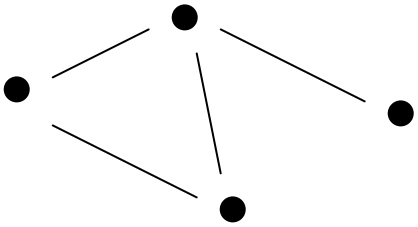


Figura 54 – exemplo de grafo conexo

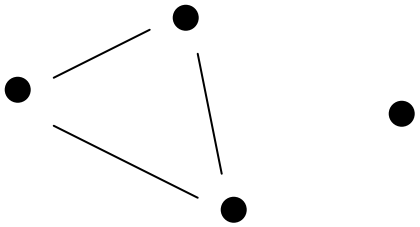


Figura 55 – exemplo de grafo desconexo

Se somente conseguirmos unir os grafos que representam os significados das palavras apenas nas sentenças, já que o fazemos pela sintaxe, e não conseguirmos unir os grafos que representam os significados das frases, então teremos vários grafos desconexos.

E ainda existe um problema decorrente desse fato: este modelo define como agramaticais sentenças que formem grafos desconexos. Isso porque, como se sabe, toda palavra existe numa sentença permitida numa língua porque há uma regra gramatical que justifique sua presença. Neste modelo, isso quer dizer que uma palavra somente pode existir numa sentença se ela participar de uma relação argumentativa. Uma palavra pode participar de uma relação argumentativa de três formas: i) como palavra regente, ii) como argumento, e iii) como intermédio de uma relação argumentativa.

Se uma palavra não participar de uma relação argumentativa, seus conceitos não se unirão aos conceitos das outras palavras da sentença. Portanto teremos um grafo desconexo. Sempre que tivermos um grafo desconexo, teremos uma construção agramatical.

Se admitirmos, entretanto, que existam grafos desconexos em relação ao texto, não podemos impedi-los em relação às sentenças. Vejam-se os textos “Comi a banana que comprei de manhã” e “Comprei bananas de manhã. Eu as comi”. Esses textos são paráfrases. Se não conseguirmos unir num único grafo as estruturas de conceitos das duas sentenças do segundo exemplo, teremos um grafo desconexo. Mas os textos são paráfrases, então os grafos que representem seus significados deverão ser idênticos. Se são idênticos, o grafo que representa os significados do primeiro exemplo também deverá ser desconexo. Mas isso não ocorre porque essa é uma sentença bem formada.

Desta forma, o modelo deverá ser capaz de unir corretamente as sentenças de um texto, a fim de conseguir representar da mesma maneira o máximo de construções com o mesmo significado. Por isso esse modelo também tratará não só a semântica e a sintaxe, mas também tratará uma parcela textual da gramática.

O que importa em relação à gramática textual é a união das diferentes sentenças no texto. Para isso basta, portanto, tratar um único fenômeno, a coesão textual.

Como sabemos, a coesão textual se dá através de coordenações, elipses, pronomes, anáforas e repetições. Da mesma forma a coesão textual se dará neste modelo.

Toda sentença que não comece por conectivo será considerada coordenada à sentença anterior. As sentenças que começarem por conectivo podem ser coordenadas a toda sentença anterior ou a apenas uma parte da sentença anterior. Para isso, deve-se tentar prosseguir na análise sintática da sentença anterior como se não houvesse o ponto a interrompendo. Se for possível analisar dessa forma a sentença, a sentença seguinte é parte da sentença anterior. É o caso do texto: “Romário tentou marcar. E conseguiu”. A sentença “E marcou” não está coordenada à sentença “Romário tentou marcar”, na verdade, o verbo “conseguiu” está coordenado ao verbo “tentou”. De qualquer forma, estes casos de coordenação e aquelas coordenações presentes dentro das sentenças, também servem para unir os grafos. No exemplo apresentado, a estrutura de conceitos do verbo “conseguir” será unida à estrutura de conceitos da sentença anterior por esse verbo se ligar a “Romário”, por coordenação com o verbo “tentou”.

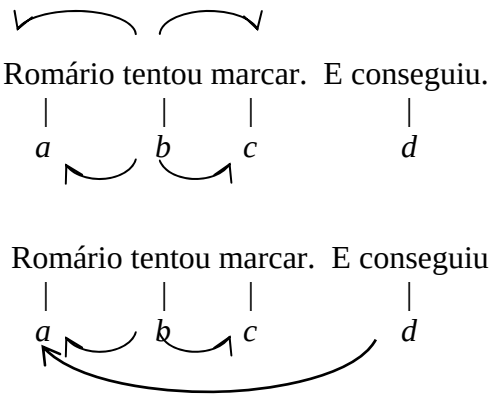


Figura 56 – coordenação 1

Nesse exemplo, também já se pode verificar um caso de união das estruturas de conceitos através de uma elipse: o verbo “conseguir” também se une ao subgrafo correspondente à estrutura de conceitos de “marcar”.

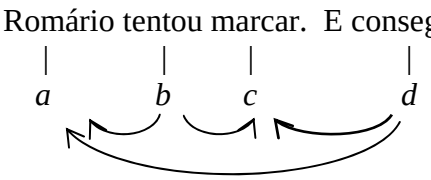


Figura 57 – coordenação 2

Os pronomes aparecem nas sentenças em lugar de substantivos, então podemos unir duas sentenças através desse substantivo referido pelo pronome.

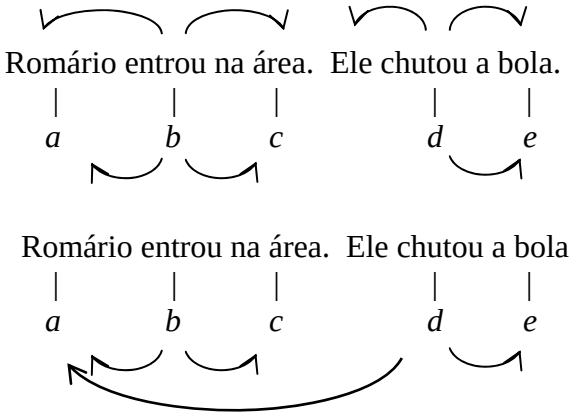


Figura 58 – anáfora 1

A anáfora ocorre quando um substantivo é usado como um pronome, isto é, referindo-se a um outro substantivo. Isso normalmente ocorre com hiperônimos.

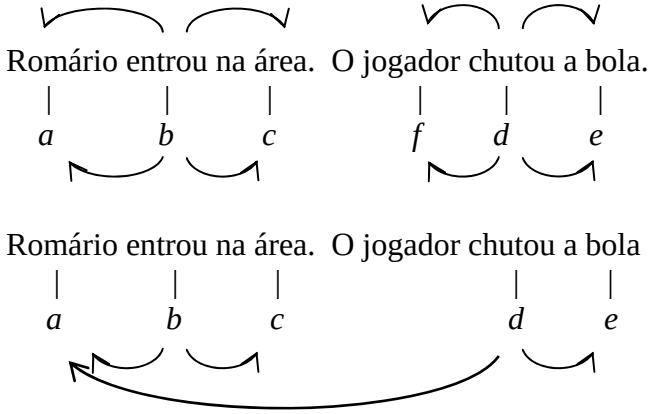


Figura 59 – anáfora 2

Ainda existe outro momento em que os grafos deverão ser unidos. Imagine um texto sobre um jogo de futebol em que apareça a palavra “Romário” várias vezes. Não precisamos repetir as estruturas de conceitos de “Romário” a todo momento em que a palavra apareça; podemos unir as estruturas de conceitos das sentenças que se uniam ao “Romário” repetido à estrutura de conceitos do primeiro “Romário” aparecido. É a isso que foi chamado de repetição.

É importante notar que o fenômeno da repetição para a coesão textual apenas acontece com substantivos. Mesmo se em um texto nos referirmos a um evento ou propriedade já mencionados, estes deverão ser retomados em suas formas substantivas. Por exemplo “O gato saiu da sala. A saída foi apressada”.

Por fim, é necessário saber que a coesão textual não será formalizada no modelo das estruturas de conceitos e argumentos. Ela será abordada em trabalhos futuros.

A coesão textual será tratada durante a análise sintática dos requisitos dos usuários e durante a análise sintática dos textos cujas buscas recairão. Desta maneira, a coesão não será formalizada em grafos, mas estará presente no algoritmo da análise sintática.

Terminada a exposição da formalização do modelo para o tratamento dos textos para a busca orientada a idéia, tentarei agora demonstrar como o trabalho de

representar os significados de um texto por esse modelo poderia ser executado por um lingüista que construísse o dicionário eletrônico do buscador orientado a idéia. Essa demonstração tem como propósito demonstrar como esse modelo se comporta, além de testar, ainda que ligeiramente, sua validade.

4
Estudo de caso

Esse capítulo demonstrará o levantamento e a atribuição de conceitos e estruturas de conceitos que representarão os significados de entradas de um dicionário eletrônico. Conjuntamente à estrutura de conceitos dessas entradas também serão demonstradas algumas estruturas de argumentos. Será simulado aqui o trabalho manual de executar essa tarefa, para o leitor se familiarizar com o modelo. No entanto, por acreditar que o levantamento das estruturas de conceitos é algo mecânico, apenas dependendo de um cotejo entre as paráfrases, para usos reais esse trabalho deverá ser automatizado.

Com a automatização do processo de atribuir estruturas de conceitos às entradas, o trabalho do lingüista passaria a se limitar ao reconhecimento das paráfrases existentes no uso analisado para cada entrada, à atribuição das classes de palavras de cada entrada e ao levantamento dos argumentos de cada entrada. O desenho final da estrutura de conceitos, a distribuição dos conceitos entre as paráfrases, o relacionamento entre conceitos e a própria criação de conceitos são tarefas possíveis e desejáveis de serem automatizadas. Criar conceitos e estruturas de conceitos não é uma tarefa difícil, apenas é trabalhosa porque envolve muitos dados – como veremos –, no entanto, é tarefa possível para a computação.

Para que fosse simulada a atribuição de estrutura de conceitos, primeiramente foi levantado um corpus. O corpus é necessário por várias razões. A primeira razão é definir sobre qual realidade lingüística particular o dicionário recai. Isso é muito importante, uma vez que o presente modelo deverá ter maior eficácia em usos restritos. Quanto mais restrita for a realidade lingüística maior será o número de paráfrases encontradas.

Isso pode ser demonstrado num exemplo simples dentro do uso escolhido para esse corpus, o jornalismo esportivo. Provavelmente, em contexto lato, não seria concebível que “contratar alguém para trabalhar num lugar” pudesse ser parafraseado por “alguém ir para um lugar”. Mas, no contexto específico tratado, isso é plenamente

concebível. Vejamos exemplos retirados de matérias de jornal que divulgam a mesma notícia:

"Falcão assina contrato com o São Paulo".

"Tricolor contrata Falcão".

"Falcão vai jogar no São Paulo".

"O ala Falcão acertou sua ida para o clube do Morumbi".

"Falcão vai para o São Paulo".

A notícia é a mesma, a informação também é a mesma, mas são usados verbos absolutamente improváveis como paráfrases. Nesse contexto específico, a informação de um clube contratar um novo jogador pode ser expressa tanto pela construção com o verbo "contratar" quanto com o verbo "ir". Isso demonstra o quanto as paráfrases são dependentes da especificidade do uso que se faz dos textos.

Essa é a primeira razão para levantar um corpus. A segunda razão é que o processo de atribuição de uma estrutura de conceitos para uma palavra é dependente de todas as paráfrases que essa palavra pode fazer no uso lingüístico trabalhado. Precisa-se do corpus para limitar quais as paráfrases reais são efetivamente realizadas nesse uso.

O corpus usado é bem reduzido para que seja possível mostrar passo a passo, no espaço de um capítulo, como executar a representação da semântica e da gramática neste modelo. O corpus é composto de oito pequenos textos retirados de matérias reais de jornais diferentes que relatam a mesma notícia: um Fla-Flu que terminou em empate. Dessas matérias foram removidas para compor o corpus apenas as frases que continham pelo menos uma destas cinco idéias: jogador, time, jogar, empatar e fazer gol[1].

Em seguida serão apresentados os processos de levantar e atribuir os conceitos das palavras de um corpus.

[1] O corpus completo está no anexo no final dessa dissertação.

4. 1
O dicionário eletrônico

O processo básico de levantar as estruturas de conceitos deve levar em conta somente as paráfrases apresentadas no corpus, como se existissem apenas as formas ali apresentadas para expressar as informações[2]. Mas se tentou, na medida do possível, ao levantar as estruturas de conceitos, prever o máximo de construções semelhantes que poderiam aparecer no contexto do jornalismo esportivo. Isso fez com que, apesar de pequenos, os exemplos de atribuição de estruturas de conceitos demonstrassem uma complexidade lingüística (e de certo modo pragmática) próxima à real.

O primeiro passo é definir, dentre as cinco idéias expressas, qual é a mais simples, isso é, qual das idéias não é composta por nenhuma outra idéia expressa.

A informação contida na expressão "fazer gol"[3] pressupõe um jogador ou um time que a realize. O mesmo se pode dizer de "jogar" e "empatar". É difícil definir entre "jogador" e "time" qual é a idéia mais simples, pois parecem se relacionar uma com a outra. Portanto, vamos trabalhar com as duas.

Sabemos pragmaticamente que um time é um conjunto de jogadores, no entanto o processo de atribuição de uma estrutura de conceitos é algo mecânico, e não exatamente pragmático. Para saber se entre as palavras ou expressões que contenham as idéias de "time" e de "jogador" há compartilhamento de conceitos, precisamos saber como essas palavras ou expressões se comportam como paráfrases.

No corpus, temos as seguintes palavras e expressões que expressam a idéia de "time", ou em que, no que expressam, está contida a idéia de time: "equipe da Gávea", "equipe flamenguista", "ambos os lados", "as duas equipes", "derrota rubro-negra", "tricolor", "equipe", "equipe tricolor", "Fla", "Fla-Flu", "Flamengo", "Flu",

[2] Por ser um trabalho inadequado às capacidades humanas, é claro que, num uso lingüístico real, as paráfrases não devem ser retiradas de um corpus uma a uma. O julgamento da presença ou ausência de uma paráfrase (bem como de um significado de uma palavra) numa realidade de uso da língua deve ser feito pelo conhecimento lingüístico de quem se propuser a representar as paráfrases. O grau de acerto do buscador orientado a idéia, portanto, será determinado pela falibilidade humana, é o preço a pagar.

[3] Como um significado se define a partir das paráfrases que podem expressá-lo, sempre que nos referirmos a uma idéia, informação etc. o faremos através de uma palavra ou de uma expressão usadas para a referirem. Isso quer dizer que quando for dito algo como "a idéia de 'time'", leia-se "a idéia da palavra 'time'". Dessa maneira, o significado também virá escrito entre aspas.

“Fluminense”, “gol tricolor”, “clube da Gávea”, “rubro-negro”, “time da Gávea”, “jogadores do Flamengo”, “tricolores”, “time das Laranjeiras”, “time rubro-negro”, “time tricolor”, “time”.

A idéia de “jogador” está contida em: “defesa rubro-negra”, “zaga rubro-negra”, “André”, “atacante”, “atacante rubro-negro”, “atleta”, “centroavante”, “Diego”, “dois jogadores rubros-negros”, “Fellype Gabriel”, “goleiro Kleber”, “Jônatas”, “Leandro”, “Lino”, “goleiro tricolor”, “gringo”, “jogador”, “meia”, “Obina”, “Pet”, “Petkovic”, “Preto”, “Preto Casagrande”, “Renato”, “Rodrigo Tiuí”, “Romeu”, “sérvio”, “Souza”, “Tuta”.

Prosseguindo, devemos emparelhar, para cada idéia, as palavras que estabeleçam algum tipo de relacionamento de paráfrases. São esses os tipos de relacionamentos: palavras ou expressões que se substituam, palavras ou expressões que compartilhem parte do significado e palavras ou expressões que contenham todo o significado de outras palavras ou expressões. Dessa maneira podemos listar:

– Palavras ou expressões que se substituam: “equipe da Gávea”, “equipe flamenguista”, “Fla”, “Flamengo”, “clube da Gávea”, “rubro-negro”, “time da Gávea”, “jogadores do Flamengo”, “time rubro-negro”; “tricolor”, “equipe tricolor”, “Flu”, “Fluminense”, “tricolores”, “time das Laranjeiras”, “time tricolor”; “ambos os lados”, “as duas equipes”; “equipe”, “time”.

– Palavras ou expressões que contenham todo o significado de outras palavras ou expressões: a idéia expressa em “equipe” e “time” está contida em “equipe da Gávea”, “equipe flamenguista”, “Fla”, “Flamengo”, “clube da Gávea”, “rubro-negro”, “time da Gávea”, “jogadores do Flamengo”, “time rubro-negro”, “tricolor”, “equipe tricolor”, “Flu”, “Fluminense”, “tricolores”, “time das Laranjeiras”, “time tricolor”.

– Todas as palavras e expressões compartilham parte do significado, o significado que é expresso por “time”.

A partir dessas divisões vamos definir as estruturas de conceitos. Se “equipe” e “time” são menos complexos que “Flamengo”, “Fluminense” etc., então devemos

começar por essas palavras. Dessa maneira, atribuiremos a elas o mesmo conceito, representando o significado de ambas as palavras[4]:

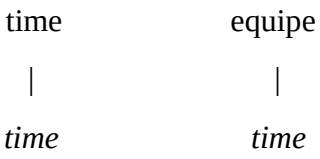


Figura 1

A palavra “Flamengo” expressa a mesma idéia de “time”, no entanto, essa idéia é especificada. Isso quer dizer que o conceito que representa o significado de “time” deve ser especificado por outro conceito ligado a ele. O mesmo ocorre com “Fluminense”, mas “Flamengo” e “Fluminense” não são especificados da mesma maneira, pois não são paráfrases, portanto os conceitos que especificarão o conceito *time* deverão ser distintos.

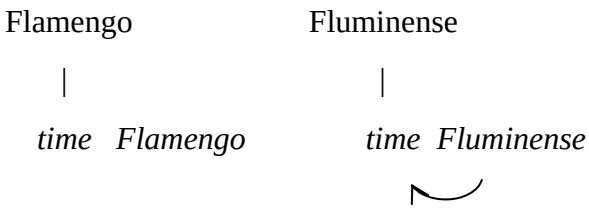


Figura 2

Nosso processo sempre começa com as palavras que representam idéias mais simples, isso é, que contenham menos idéias que as outras palavras analisadas, e sempre escolhemos começar pelas palavras simples e não pelas expressões

[4] É muito importante notar que a escolha das palavras usadas para nomear um conceito é arbitrária, poderia ser uma letra, como foi feito anteriormente, poderia ser uma seqüência aleatória de caracteres ou números, poderia ser um símbolo, um desenho ou uma cor, poderia ser ainda palavras sem nenhuma ligação com os significados (como o conceito *chuchu*, representando a idéia de “time”). O que importa é que sempre que se esteja referindo a uma determinada idéia, esta seja representada pelo mesmo conceito ou estrutura de conceitos. Escolheu-se para nomear os conceitos palavras que tivessem alguma ligação com os significados por uma necessidade mnemônica. Para diferenciar o conceito *time* da palavra “time”, o conceito vem escrito em itálico e a palavra entre aspas. Escolhemos arbitrariamente o conceito *time* para representar a idéia expressa pela palavra “time”, segundo nossos critérios, poderia ser *equipe*, por exemplo. Mas, após feita a escolha por *time*, todas expressões que contenham essa idéia devem possuir em suas estruturas de conceitos o conceito *time* e não outro.

compostas. Assim sendo, as próximas palavras a receber estruturas de conceitos são os epítetos “tricolor” e “rubro-negro”. Essas palavras agem como sinônimos respectivamente de “Fluminense” e “Flamengo”. Então receberão a mesma estrutura de conceitos que essas palavras.

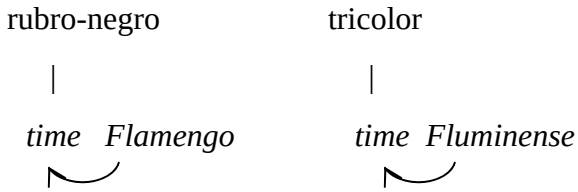


Figura 3

O mesmo acontece com “Fla” e “Flu”:

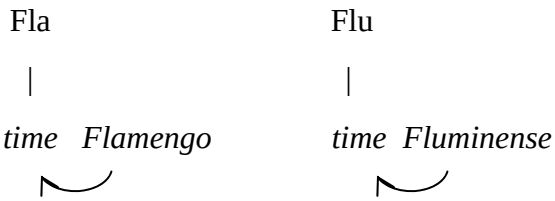


Figura 4

É muito interessante notar as peculiaridades geradas pelas premissas desse modelo. Como um conceito se relaciona a outro pela especificação, acabamos não tendo uma estrutura hierarquizada entre os conceitos *time* e *Fluminense*, como uma interpretação pragmática calmamente estratificaria. Neste modelo, nunca poderá existir hierarquia entre os conceitos por dois motivos. O primeiro é que se quer aproveitar as habilidades do grafo[5] (como a busca de subgrafos em grafos) no nosso modelo, e num grafo não há hierarquia entre seus elementos. A árvore, uma técnica

[5] Lembre-se que numa ciência aplicada, como é o caso da Lingüística Computacional, as técnicas como grafos, árvores, funções etc. são usadas para representar um fenômeno da realidade. Não podemos usar qualquer técnica para representar qualquer fenômeno da realidade. Pode-se representar os significados por grafos porque os significados, conforme se está analisando-os, possuem certas propriedades que os permitem serem representados por grafos. Porém pode haver mais de uma técnica passível de ser usada para representar um fenômeno estudado. A escolha de qual técnica usar deve ser feita tendo em mente as habilidades mais adequadas para nossos fins que essas técnicas possuem.

próxima ao grafo que guarda a propriedade da hierarquia entre seus elementos, não possui algumas habilidades de um grafo úteis para busca orientada a idéia[6].

O segundo motivo de não hierarquizar os conceitos é que estamos utilizando apenas um único critério que relaciona os conceitos: a especificação. Se, em nosso modelo, um conceito está ligado a outro por uma seta, isso significa que um conceito está especificando o outro. Por um acaso, sabemos pragmaticamente que entre o conceito de um hipônimo e os conceitos de um hiperônimo (como é o caso de "Flamengo" e "time") há uma relação do tipo *é um*. Essa é uma relação hierárquica. Mas entre outros conceitos a relação pragmática pode não ser hierárquica. Por exemplo, na expressão "gato gordo", um conceito da estrutura de conceitos de "gordo" terá que se ligar por especificidade (já que a relação argumental especifica os significados das palavras) a um conceito da estrutura de conceitos de "gato". É muito provável que não possamos admitir que a relação pragmática existente entre esses conceitos seja passível de uma hierarquização. Discutiremos mais esse ponto no capítulo 2, mais exatamente quando se discutir sobre redes semânticas e *frames*.

Agora podemos passar para as expressões compostas. A análise semântica das expressões compostas necessita de alguma análise gramatical. Temos as expressões "equipe da Gávea", "clube da Gávea", "time da Gávea", "equipe flamenguista" e "time rubro-negro", expressando a mesma idéia expressa por "Flamengo", e as expressões "equipe tricolor", "time tricolor" e "time das Laranjeiras", expressando a mesma idéia expressa por "Fluminense".

Entre "equipe da Gávea", "clube da Gávea" e "time da Gávea" apenas as palavras "equipe", "clube" e "time" se alternam, e todas essas palavras detêm o conteúdo significativo representado pelo conceito *time*. Dessa maneira, podemos usar a estrutura de argumentos para definir a estrutura de conceitos. Podemos considerar que a palavra "Gávea" pede um complemento (um substantivo) que possua o conceito

[6] A saber, a habilidade não encontrada na árvore que mais impossibilita o recurso a essa técnica de representação para nossa finalidade é a incapacidade dela de ignorar a seqüência em que seus elementos estão ordenados. Devido à hierarquização de seus elementos, a ordem em que eles aparecem na árvore é fundamental. Isso impediria certas habilidades como a busca de sub-estruturas (como os subgrafos) dentro de estruturas maiores em que se ignorem elementos que aparecem intercalados nessa estrutura ou que se ignore a ordem em que os elementos se dispõem na estrutura.

time em sua estrutura de conceitos, sendo que a ligação entre esse complemento e a palavra “Gávea” é mediada pelo artigo “a” e a preposição “de”. Desta maneira:

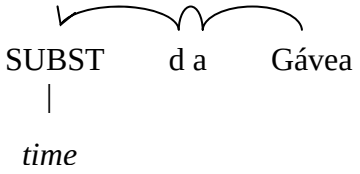


Figura 5

Se o conjunto “equipe/time/clube da Gávea” tem o mesmo significado de “Flamengo”, então esse conjunto gera uma estrutura de conceitos final com o esquema *time Flamengo*. E se é para as palavras “equipe”, “time” e “clube” que é distribuído o conceito *time*, então resta apenas o conceito *Flamengo* para “Gávea”. No entanto, não é apenas esse conceito que forma a estrutura de conceitos de “Gávea”, isso porque o conceito pedido a um complemento e a seta que representa a relação de especificação fazem parte da estrutura de conceitos de uma palavra. Assim, esta será a estrutura de conceitos de Gávea:

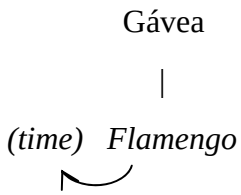


Figura 6

O mesmo acontece com “time das Laranjeiras” e seus correlatos.

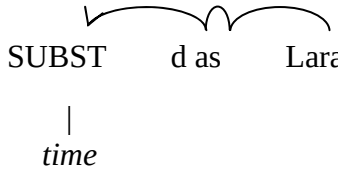
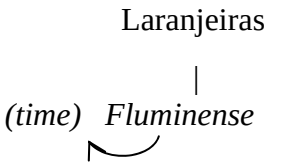
Estrutura de argumentos	Estrutura de conceitos
	

Tabela 1

No caso de “time rubro-negro” e “equipe tricolor”, o processo é semelhante. A única diferença está na estrutura de argumentos de “rubro-negro” e de “tricolor” que não pedem ligação mediada.

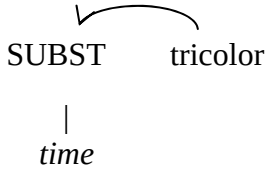
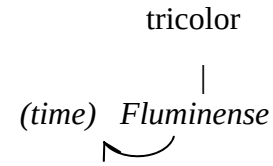
Estrutura de argumentos	Estrutura de conceitos
	

Tabela 2

Mas esse é apenas um dos significados de “tricolor” e “rubro-negro”. Há algumas páginas atrás tinha sido apresentada outra estrutura de conceitos para essas palavras, a mesma estrutura de conceitos de “Fluminense” e “Flamengo” respectivamente. São esses sentidos que prevalecem em situações como as das construções “derrota rubro-negra” e “gol tricolor”, pois podem ser parafraseadas por “derrota do Flamengo” e “gol do Fluminense”.

Podemos encerrar por enquanto esse grupo de palavras e passarmos para as palavras que compartilham o sentido de “jogador”. A primeira questão a ser levantada é a respeito da proximidade morfológica entre “jogador”, “jogar”, “jogo”, “jogado” e “jogada”. Essas palavras compartilham estruturas de conceitos?

Vamos tratar por enquanto apenas de “jogador”. No nosso contexto específico, não devemos tratar a relação entre “jogador” e “jogar” como podemos tratar a relação de paráfrases existente entre, por exemplo, “destruir” e “destruidor” das sentenças “José destruiu a casa” e “José foi o destruidor da casa”. No nosso contexto, “jogador” é apenas uma categoria de pessoa, é um tipo específico de pessoa, como o são os técnicos, os árbitros e os torcedores. Isso porque não vamos encontrar uma paráfrase para “jogar”, usando “jogador” em seu lugar.

Então, por ora, não precisaremos nos preocupar com a estrutura de conceitos de palavras como “jogar” para definir a estrutura de conceitos de “jogador”. Mas, como havia sido previsto em outro momento, “jogador” estabelece uma relação com “time”.

Pragmaticamente sabemos que um time é um conjunto de jogadores e que um jogador é um elemento de um time. No entanto, como não foi encontrada nenhuma paráfrase para "time" que contivesse a palavra "jogador" substituindo-a ou substituindo parte de seu significado, então excluímos o significado de "jogador" das palavras que continham o significado de "time".

A relação entre essas palavras acontece na estrutura argumental da palavra "jogador", que pede como complemento substantivos que contenham *time*, como ocorre em sentenças como "o jogador do Flamengo". Esse complemento ainda pode ser um adjetivo-classificador (adjetivo derivado de substantivo), como em "o jogador flamenguista" ou "o jogador rubro-negro".

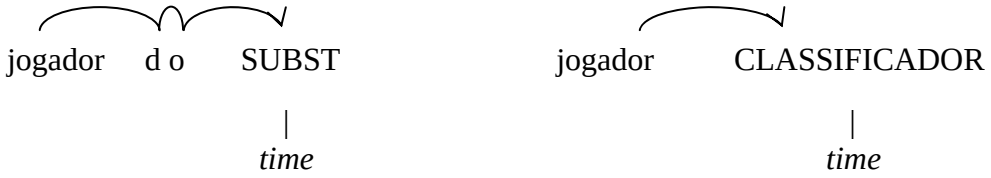


Figura 7

Com respeito à estrutura de conceitos propriamente dita de "jogador", podemos seguir passos semelhantes aos seguidos anteriormente e teremos:

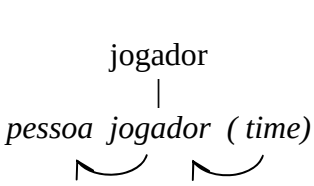
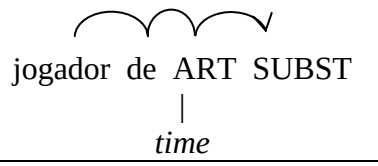
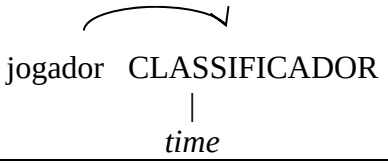
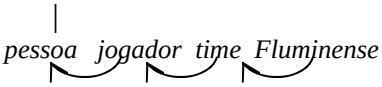
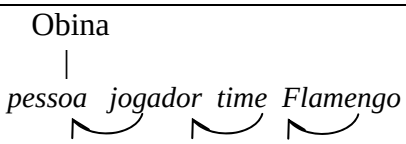
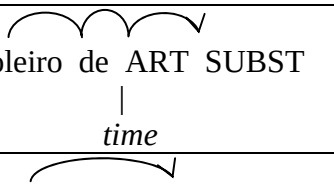
PALAVRA	ESTRUTURA DE CONCEITOS	ESTRUTURA DE ARGUMENTOS
jogador		
		
Petkovic		Petkovic
Obina		Obina
goleiro		
		
Kleber		Kleber

Tabela 3

Na realidade, quando me referi ao fato de não encontrarmos nenhum exemplo de paráfrases entre “jogador” e “time”, soneguei um caso em particular. Existe um momento em que podemos substituir “jogador” por “time”: as construções em que exista a idéia de “fazer gol”. Por exemplo, sabendo-se que Petkovic é um jogador do Fluminense, podemos parafrasear a frase “Petkovic marcou um gol” por “o

Fluminense marcou um gol". Mas o contrário não pode ser feito. Não podemos substituir "o Fluminense marcou um gol" por "Petkovic marcou um gol", nem por "um jogador do Fluminense marcou um gol", nem mesmo por "um jogador marcou um gol".

A sentença "o Fluminense marcou um gol" não pode ser parafraseada por "Petkovic marcou um gol" porque não podemos atribuir o gol a Petkovic somente a partir da primeira sentença. O mesmo ocorre com o veto à paráfrase com "um jogador do Fluminense marcou um gol"; não podemos prever que o gol tenha sido feito por um jogador do Fluminense, pois há gols contra. Já o caso da paráfrase com "um jogador marcou um gol" é vetado porque, apesar de a sentença "o Fluminense marcou um gol" pressupor pragmaticamente que algum jogador tenha marcado o gol (afinal somente jogadores podem marcar gols), não podemos parafraseá-las simplesmente porque uma frase como "um jogador marcou um gol" sem qualquer especificação sobre esse gol não ocorre nesse contexto.

Informações óbvias demais não aparecem num jornal. Se uma sentença é escrita para informar sobre um gol, isso vem sempre acompanhado da informação sobre quem o marcou ("gols de Tuta e Diego Souza"), ou sobre que time o marcou ("o Fluminense quase marcou o gol"), ou sobre o modo como o gol foi marcado ("Pet fez um gol de placa") etc.

Se uma paráfrase não existe num contexto simplesmente porque uma frase potencial nunca ocorre, então não precisamos nos preocupar em prevê-la.

Somente ocorre paráfrase entre uma sentença que informe a qual time pertence o jogador que marcou o gol e uma sentença que informe o nome do time que marcou o gol. Desse modo, a idéia de "Fluminense marcou um gol" está contida em "um jogador do Fluminense marcou um gol" e em "Petkovic marcou um gol". O termo "Petkovic" possui em sua estrutura de conceitos a informação de que este é um jogador do Fluminense, como mostramos há algumas páginas.

Esse tipo de paráfrase só ocorre com a informação "fazer gol". Uma frase como "Petkovic se machucou" não pode ser parafraseada por "o Fluminense se machucou". Essa foi a razão por termos ocultado esse caso de paráfrase entre "jogador" e "time" durante a definição das estruturas de conceitos para as palavras que contivessem essas

idéias. A paráfrase encontrada tem que ser prevista na estrutura de conceitos de palavras que contenham a idéia de “marcar gol”, ficando assim inalteradas as estruturas de conceitos de palavras com a idéia de “jogador” e de “time”.

Então começaremos a trabalhar a estrutura de conceitos de palavras ou expressões que contenham a idéia de “marcar gol”. A solução encontrada para a estrutura de conceitos de “marcar gol” incluir a paráfrase encontrada se fez por acréscimos de conceitos e por uma distribuição de ligações entre esses conceitos sem nenhum correspondente a interpretações do significado de “marcar gol”, “jogador” ou “time”. Foi um ato simplesmente mecânico, e por isso vou apresentá-lo de forma mecânica, isso é, com o uso de conceitos representados novamente por letras. Depois as substituiremos por formas mais mnemônicas.

Atribuiremos a cada palavra um conceito e os ligaremos conforme a sintaxe.

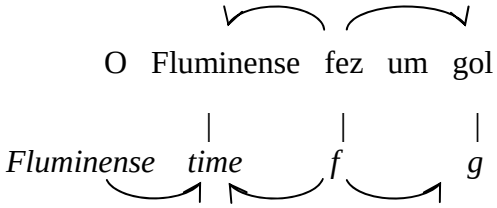


Figura 8

E o mesmo faremos com a construção com a idéia de “jogador”.

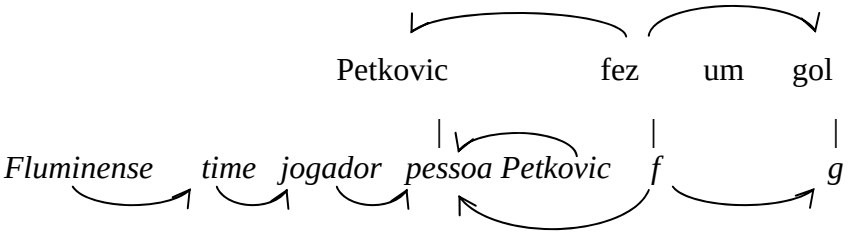


Figura 9

Por enquanto, as duas construções não formam paráfrases. Note-se que, da maneira como foram formalizadas, o significado da primeira sentença não está contido na segunda sentença. Para isso acontecer, o grafo que representa a junção das estruturas de conceitos das palavras da primeira sentença deveria ser um subgrafo da

estrutura de conceitos da segunda sentença. Isso não está ocorrendo porque o conceito *f* da primeira sentença está ligado a *time* e na segunda sentença está ligado a *pessoa*.

Para resolvermos isso, as estruturas de conceitos de “fez um gol” têm de variar conforme incidam sobre “time” ou sobre “jogador”. Isso significa dizer que a estrutura de conceitos muda conforme a estrutura de argumentos mude.

O primeiro passo para isso é alterar a estrutura argumental de “fazer um gol”. O verbo “fazer” é um verbo que se comporta como “dar” em “dar um abraço”, que na verdade é uma forma analítica de expressar “abraçar”. Em casos como esse, o verbo age como uma palavra de ligação, e não precisa gerar estrutura de conceitos. Veja:

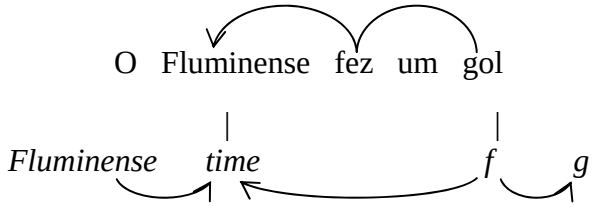


Figura 10

Dessa maneira os conceitos antes atribuídos a “fez” serão agora atribuídos a “gol”. A palavra “gol” também terá de mudar de classe, sendo agora um substantivo-evento. Somente substantivos-evento participam dessa forma de construção. Devemos notar o fato de *f* passar a receber a mesma ligação que liga “gol” a seu complemento.

O conceito *f* se liga ao conceito *pessoa* de “Petkovic”, impossibilitando a paráfrase. Por esse mesmo processo que é impossibilitado que frases como “Petkovic caiu” e “o Fluminense caiu” não fossem paráfrases, isso é, o as ordenações das relações nos grafos impossibilita que a estrutura de conceitos formado pela sentença “o Fluminense caiu” não contenha a mesma estrutura de conceitos gerada por “Petkovic caiu”.

Por enquanto sabemos que uma das entradas de “gol” terá a estrutura de conceitos e a estrutura de argumentos seguintes:

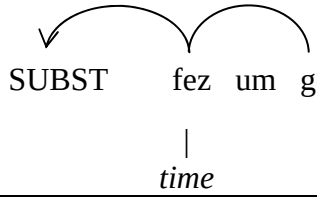
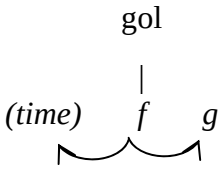
Estrutura de argumentos	Estrutura de conceitos
	

Tabela 4

Outra deverá ser a entrada de “gol”, já que pretendemos modificar sua estrutura de conceitos. Lembre-se que diferentes sentidos determinam diferentes entradas. Nessa entrada acrescentaremos um novo conceito *d* em “gol”. Ele se ligará ao conceito *pessoa* do complemento. O conceito *f* continuará na estrutura, ligando-se ao conceito *time* do complemento. Veja:

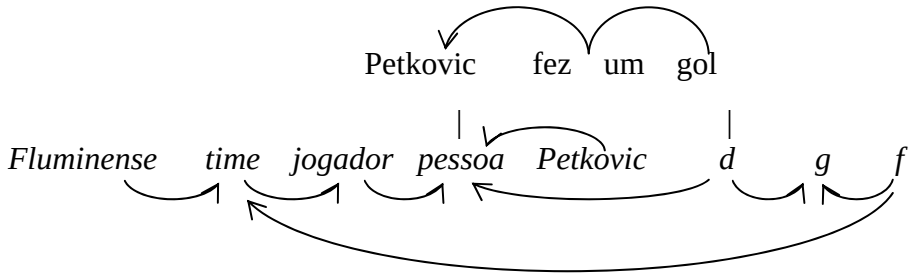


Figura 11

Podemos substituir por formas mais mnemônicas esses conceitos representados por letras. Escolhi substituir *d* por *fazer*, *g* por *gol* e *f* por *possuir*. Essa escolha foi completamente arbitrária, como, aliás, também o foi a escolha de todos os demais nomes de conceitos. As estruturas de conceitos e de argumentos de “gol” ficam assim:

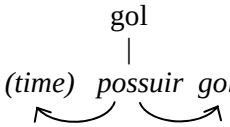
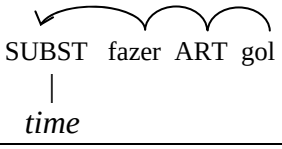
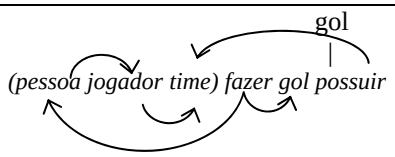
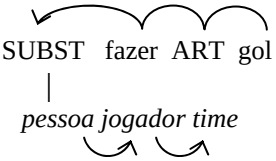
PALAVRA	ESTRUTURA DE CONCEITOS	ESTRUTURA DE ARGUMENTOS
gol		
gol		

Tabela 5

As mesmas estruturas de conceitos são atribuídas a outras palavras que contenham a idéia de “marcar gol” e outras ainda podem ser as estruturas de argumentos de “gol”. Por exemplo, em sentenças como “o gol de Petkovic”.

Passemos agora para a idéia de “jogar”. Novamente devemos saber se em palavras como “jogo”, “jogada” e “jogador” a idéia de “jogar” está presente. Como foi dito antes, “jogador” não contém a idéia de “jogar”, isso porque, se admitíssemos que o significado de “jogar” estivesse contido em “jogador”, toda sentença em que aparecesse a palavra “jogador” conteria a idéia de “jogar”, mesmo em sentenças como “o jogador treinou a manhã inteira” ou, pior, “Petkovic concedeu uma entrevista coletiva”. Mas talvez o oposto seja verdade.

O significado de “jogador” pode estar inserido no significado de jogar, como se “jogar” significasse algo como “ser jogador”. Isso poderia ser usado para permitir paráfrases do tipo que substituísse “Petkovic foi o melhor jogador em campo” por “Petkovic jogou melhor que todos em campo”. Ou talvez esse fosse um caso em que “jogador” possuísse um outro significado, o significado de “jogar”.

É muito comum, no momento em que estamos definindo as estruturas de conceitos das paráfrases, nos debatermos com casos em que mais de uma distribuição é possível, e precisarmos fazer uma escolha. Não existe certo ou errado nas definições de estruturas de conceitos. Existem escolhas que funcionam e outras que não funcionam. E “funcionar” significa tornar o computador capaz de reconhecer

sentenças com as mesmas informações como sendo paráfrases. Isto é, existem escolhas melhores e escolhas piores.

Escolhas que façam com que o computador trabalhe com menos dados serão melhores do que as que exijam mais poder computacional. Portanto, sempre são melhores as distribuições que possuem menos conceitos. Também sempre são melhores as distribuições que geram menos homonímias. E também são melhores as distribuições que conseguem tratar o maior número de paráfrases de uma só vez. O problema é que esses três fatores são conflitantes e muitas vezes conciliá-los é impossível.

Particularmente prefiro tratar o maior número de paráfrases de uma só vez, mesmo que isso gere muitas homonímias e muitos conceitos. Mas isso é uma escolha pessoal, não uma obrigação. Essa alternativa talvez nem seja a mais recomendável para o trabalho em equipe, pois muitas vezes uma representação com muitos conceitos pode ficar bem confusa.

Não precisamos nos preocupar nesse estudo de caso com os eventos em que "jogador" e "jogar" serão paráfrases, já que esses casos não aparecem em nosso corpus. Mas não podemos dispensar as paráfrases entre "jogar" e "jogo".

Os momentos em que aparecem as palavras "jogo", "partida", "clássico" e outras palavras ou expressões semelhantes, elas informam: a) o andamento do jogo ("o jogo melhorou muito"), b) o local onde o jogo ocorreu ("o clássico foi em Volta Redonda"), c) se o jogo terminou ("o fim do jogo"). Em todos esses casos as palavras ou expressões com a idéia de "jogo" devem se relacionar com a idéia de "jogar". Poderíamos parafrasear os exemplos assim: "jogou-se melhor", "os times jogaram em Volta Redonda", "acabou-se de jogar".

Se quisermos acrescentar outros casos possíveis no uso tratado em que "jogar" e "jogo" formassem paráfrases, poderíamos citar o caso de: "o Flamengo fez um bom jogo" e "o Flamengo jogou bem". E ainda complicaríamos mais a distribuição de estruturas de conceitos ao parafrasear as construções também por "o Flamengo fez uma boa atuação".

E ainda existe, nesse corpus, um momento em que “jogo” não compartilha idéias com “jogar”. Isso acontece em casos como “o jogo empatou” e “a partida virou”.

Vamos tratar, portanto, um caso de cada vez. Primeiramente trataremos a estrutura de conceitos de “jogar”, já que ela sempre será parafraseada por “jogo”, enquanto o contrário não ocorre. O primeiro fato a ser notado é que existe uma diferença pragmática entre “time jogou” e “jogador jogou”, no que diz respeito à compreensão que fazemos de “jogar” em cada um dos casos.

Devemos ter cuidado com a compreensão que temos das palavras. Muitas vezes elas nos dão pistas úteis para a composição das estruturas de conceitos, outras vezes, no entanto, nos levam a caminhos enganadores. Nesse caso, a compreensão parece ser uma pista segura, uma vez que ela se relaciona com paráfrases. Em sentenças compostas pela idéia de “o time jogou”, podemos nos referir à atuação do time (“o Flamengo jogou bem”) ou a qual foi o adversário do time (“o Flamengo jogou contra o Fluminense”). Já em sentenças que se componham de “o jogador jogou”, as informações podem informar sobre a atuação do jogador (“Petkovic jogou mal”) ou sobre o fato de o jogador ter sido ou não escalado para o jogo (“Felipe não jogou a partida”). Vamos ignorar somente esse último caso de “jogador jogou” porque não aparece em nosso corpus.

De forma a apresentar o máximo de técnicas de composição de estruturas de conceitos, vamos usar outra estratégia aqui. Usaremos o mesmo conceito para “jogar” em todos os casos apontados. O que vai vetar as paráfrases serão as estruturas de argumento possíveis de “jogar”. Vejamos:

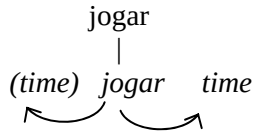
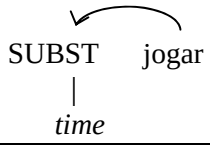
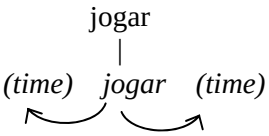
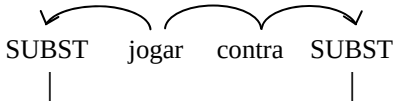
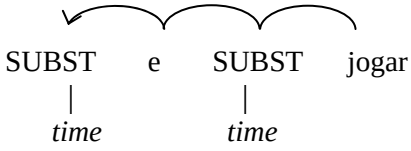
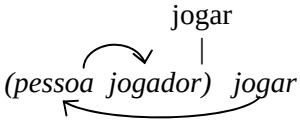
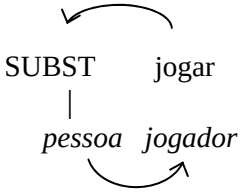
PALAVRA	ESTRUTURA DE CONCEITOS	ESTRUTURA DE ARGUMENTOS
jogar		
jogar		 
jogar		

Tabela 6

Essas mesmas estruturas de conceitos podem ser repetidas nas palavras sinônimas “jogo” e “partida” que se comportarão como paráfrases em todos os casos de jogar .

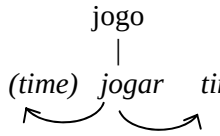
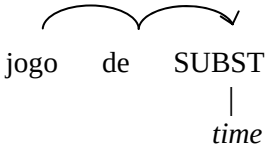
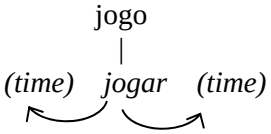
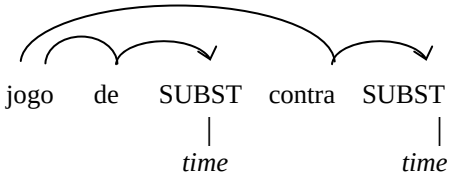
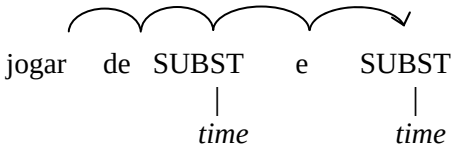
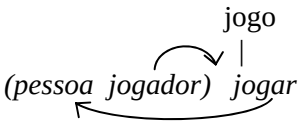
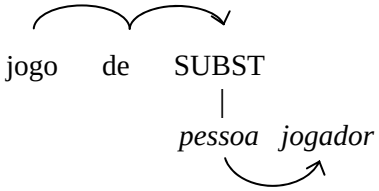
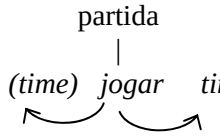
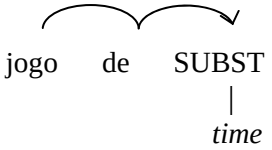
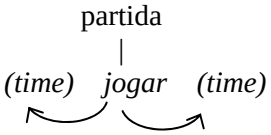
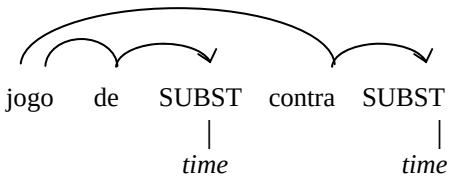
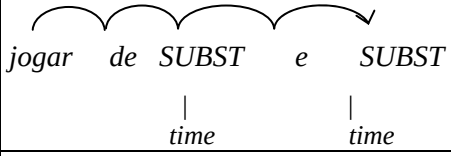
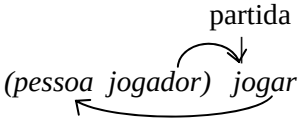
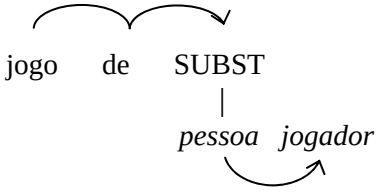
PALAVRA	ESTRUTURA DE CONCEITOS	ESTRUTURA DE ARGUMENTOS
jogo		
jogo		 
jogo		
partida		
partida		 
partida		

Tabela 7

Para os outros sentidos de “jogo” e “partida” optamos pelo conceito *partida*.

PALAVRA	ESTRUTURA DE CONCEITOS	ESTRUTURA DE ARGUMENTOS
jogo	jogo | *partida*	jogo
partida	partida | *partida*	partida

Tabela 8

Esses sentidos de “jogo” e “partida” são usados nos casos “a partida/o jogo virou” ou “a partida/o jogo empatou”. Nesses casos, tanto “virou” quanto “empatou” possuem a idéia de “marcar gol” da forma como já havíamos apresentado.

A idéia de “empatar” é uma das que fazem parte de nosso corpus. Nele, encontramos dois usos para “empatar”. O primeiro uso é caracterizado por “empatar” contendo a idéia de “fazer gol”. Vejam-se os exemplos:

“O empate do Flamengo aconteceu”.

“O time da Gávea conseguiu empatar”.

“Renato empatou a partida”.

“O Flamengo não desistiu e chegou ao empate”.

Nesses casos, poderíamos parafrasear as sentenças com palavras que contivessem a idéia de “marcar gol”.

“O gol do Flamengo aconteceu”.

“O time da Gávea conseguiu marcar”.

“Renato fez um gol”.

“O Flamengo não desistiu e conseguiu o gol”.

É interessante notar que com as substituições não há perda de informações. A informação trazida com “empatar”, nesses casos, traz consigo a idéia “marcar gol” e também pressupõe outro evento “marcar gol”, anterior e atribuído ao time adversário em relação àquele que empatou. Em matérias como as que compõem nosso corpus, nas quais uma partida de futebol é narrada, no momento em que uma sentença como

“Renato empatou a partida” surge no texto, é certo que anteriormente o texto já havia trazido a informação sobre um gol do time adversário. Portanto, se a frase tivesse sido escrita como “Renato fez um gol”, a informação permaneceria inalterada.

No entanto, será que tratar “empatar” e “marcar gols” como paráfrases perfeitas é a melhor escolha para um buscador orientado a idéia?

Isso não seria desejável para o buscador. Imagine que um usuário queira encontrar sentenças que informem que um time fez um gol de empate. Se sempre a informação de “empatar” nesse sentido for decomposta em estrutura de conceitos do mesmo modo como “marcar gol” é decomposta, o buscador retornará para o usuário todas as sentenças em que aparece “marcar gol”, inclusive sentenças como “o Flamengo virou o jogo”. E não era isso o que o usuário esperava.

Podemos pensar na estratégia oposta. Nesse uso lingüístico, sempre que um gol influencia no resultado parcial do jogo, isso é, sempre que um gol puser um time na frente do placar, empatar ou virar a partida, isso será informado explicitamente como tal. Isso quer dizer que, se um gol empata a partida, ele sempre será informado como um gol de empate. A matéria nunca irá informar simplesmente a seqüência de gols, deixando que o leitor vá somando e deduzindo quem está com a vitória, derrota ou empate parciais.

Dessa maneira, não precisaríamos atribuir o significado de “marcar gol” para “empatar” já que a idéia de “empatar” sempre aparecerá expressa com palavras que incluem a idéia de “empatar” e nunca com palavras que somente incluem a idéia de “marcar gol”.

No entanto, acredito que seja uma boa habilidade do buscador encontrar uma sentença como “Petkovic chuta e empata a partida”, quando o usuário requisitar “Fluminense fez gol”. Se assim quisermos, “empatar” deverá conter em sua estrutura de conceitos a idéia de “fazer gol”.

A melhor estratégia então é tratarmos esse problema de forma mecânica. O gol de empate é um tipo específico de gol, que se distingue de um gol que vira a partida ou de um gol que põe o time na frente do placar. Usaremos então apenas um conceito ligado ao conceito *gol* e aproveitaremos todo o resto da estrutura de conceitos de “marcar gol”.

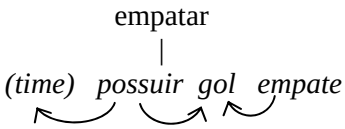
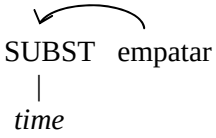
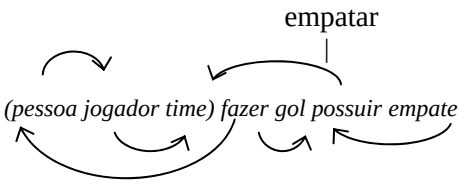
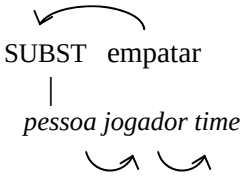
PALAVRA	ESTRUTURA DE CONCEITOS	*ESTRUTURA DE ARGUMENTOS*
empatar		
empatar		

Tabela 9

Para percebermos a diferença entre esse uso de “empatar” e o outro, será interessante notar também que palavras com a idéia de “vencer” ou “perder” não podem ser usadas nesse caso.

“*A vitória do Flamengo aconteceu”.

“*O time da Gávea conseguiu vencer”.

“*Renato venceu a partida”.

“*O Flamengo não desistiu e chegou à vitória”.

Note-se que considero como impossíveis as frases acima, considerando-as no mesmo contexto que as originais, isso é significando um momento em que um time ou um jogador fez um gol. Vejamos que, se aceitarmos tais sentenças, a compreensão que fazemos é que elas nos informam sobre um resultado final de um jogo, e, nesse caso, a terceira sentença ainda permaneceria impossível.

A idéia de “resultado final de um jogo” é o segundo uso de “empatar”. Esse caso aparece preferencialmente nas manchetes, no início e no fim das matérias. São momentos em que “vencer” e “perder” poderiam ser usados.

“Clássico Fla-Flu termina empatado”.

“Fluminense e Flamengo empataram em 2 a 2”.

“Em clássico emocionante, Fla-Flu acaba empatado”.

“Flamengo e Fluminense empatam em Volta Redonda”.

"O resultado acabou com um empate emocionante em Volta Redonda".

Note-se que esses casos não podem incluir a idéia de "marcar gol", isso porque um empate pode acontecer sem gols. A informação sobre os gols pode ser dada na estrutura argumentativa da palavra que contenha essa idéia de "empatar". É o que acontece em construções como "empate por 2 a 2", por exemplo. Esse tipo de estrutura em que aparece o placar é típica desse uso de "empatar" e não ocorre no outro uso.

Nesse uso, a estrutura de conceitos de palavras com a idéia de "empatar" não precisa incluir a estrutura de conceitos de "marcar gol". Mas é importante notar que a sentença "Flamengo e Fluminense empataram em 2 a 2" deve ser retornada pelo buscador se o usuário requisitar "Fluminense fez gol". Entretanto, o buscador não deve retornar para o mesmo requisito do usuário sentenças como "Flamengo e Fluminense empataram", nem mesmo "Flamengo e Fluminense empataram em 0 a 0". Nesse último caso, nem Flamengo nem Fluminense marcaram gols. Um antônimo ou uma negação de algo nunca podem ser retornados pelo buscador quando esse algo for requisitado pelo usuário.

4. 2
Problemas ainda não resolvidos pelo modelo

Como se viu nos exemplos de atribuição de conceitos e relações entre os conceitos apresentados, esse trabalho é bem complicado. A complicação não decorre de ser um trabalho difícil, mas por exigir a atenção para inúmeros detalhes. Quem se propuser a levantar a estrutura de conceitos de uma palavra deve prever todos os seus usos, todas as paráfrases de seus usos, todas as palavras e expressões que podem compartilhar parte de sua estrutura de conceitos em cada uso, todas as palavras e expressões que estão contidas no significado da palavra em todos os seus usos. Trabalhar com tantos dados assim numa situação real de sistemas de recuperação de informações em corpora textuais, cujo número de palavras pode chegar facilmente a dezenas de milhares, é algo que beira a incapacidade humana.

Mas é justamente a capacidade de trabalhar com milhões de dados e variáveis computáveis interagindo entre si que faz da computação a ferramenta útil que é. O

melhor seria encontrar uma forma de automatizar o processo que consiga considerar como paráfrases duas sentenças diferentes. Como já foi frisado inúmeras vezes, a atribuição das estruturas de conceitos é algo mecânico, que muitas vezes é inclusive atrapalhado pelo nosso entendimento das palavras.

O problema que impede a automatização desse processo no estágio atual de desenvolvimento do modelo de representação da semântica para a busca orientada de textos é justamente o fato de ele ainda estar incompleto.

Algumas questões permanecem em aberto, impossibilitando a automatização. Alguns problemas não resolvidos por esse modelo dizem respeito a algumas classes de palavras ainda não trabalhadas, como os advérbios, e também estruturas oracionais, principalmente as orações adverbiais, cujas conjunções devem influenciar as estruturas de conceitos e paráfrases. Por exemplo, deverá haver um conceito para indicar a idéia de causa que se pode notar em alguns eventos, sendo que essa idéia pode ser expressa também em orações adverbiais causativas.

Outros problemas dizem respeito à estrutura textual. O texto também possui uma gramática peculiar. Nos exemplos apresentados de jornalismo esportivo pode-se notar que a manchete tem uma gramática diferente do restante do texto: por exemplo, o artigo é quase que totalmente abolido nas manchetes em momentos em que, se as mesmas frases estivessem escritas no corpo do texto, seria obrigatório o uso do artigo.

A teoria terá que prever essas gramáticas concorrentes.

Em certos momentos, uma desambiguação pode ser feita simplesmente porque uma expressão aparece escrita num determinado espaço do texto. Por exemplo, foram apresentados dois sentidos para a palavra “empatar”, aquele que informava sobre um gol marcado que igualava o placar da partida ou aquele que informava o resultado final da partida. Se o verbo “empatar” estiver escrito na manchete, no início ou no final da matéria, é mais provável que seu sentido seja o de resultado de uma partida. Se o verbo estiver escrito no desenvolvimento da matéria é provável que o sentido seja o outro.

Outro problema latente diz respeito às flexões, principalmente as verbais. Provavelmente, a flexão de tempo terá de ser tratada na estrutura de conceitos,

enquanto a flexão de modo (como talvez também a flexão de número e pessoa) deverá ser tratada na estrutura de argumentos.

Os tempos verbais deverão ser simplificados como ações que ocorreram antes, ao mesmo tempo ou depois de outras ações, não importando se ocorreram no passado, presente ou futuro. Isso porque para o usuário não terá muita importância saber se um evento aparece escrito no presente, no passado ou no futuro. Mas a relação “antes”, “durante” e “depois” entre os eventos é importante para as paráfrases, como é o caso de “empatar”, que poderia ser parafraseada por dois eventos “marcar gol”, sendo que um ocorreria em momento anterior ao outro.

Talvez seja importante informar também se uma ação foi repetida, se prolongou no tempo ou se foi encerrada; essas informações são conseguidas na análise aspectual dos tempos e modos verbais. Portanto temos que teorizar como os diferentes tempos e modos verbais e a interação entre esses tempos e modos de diferentes verbos informam todos esses dados.

Quanto mais paráfrases conseguirmos tratar, mas resultados corretos nosso buscador retornará.

4. 2. 1
Trabalhos futuros

O tratamento da semântica entendida enquanto paráfrase se mostra bom para aumentar os resultados relevantes de uma busca bem como restringir os resultados pouco relevantes.

No entanto, como se trata de uma forma de representar a semântica nos arquivos de texto, outros usos podem ser pensados para este modelo. O primeiro deles é a tradução automática. Se pensarmos que uma tradução é uma paráfrase de um texto, por ser uma reescrita em outra língua, entenderemos por que esse modelo pode ser usado para essa finalidade. Para isso, entretanto, não bastaria desenvolver o dicionário eletrônico das duas línguas envolvidas nas traduções, seria necessário desenvolver um dicionário eletrônico da tradução de uma língua para outra. Isso porque as estruturas de conceitos atribuídas às palavras não são universais, são apenas daquela língua e num certo uso especifico que se faz dela. Para cada uso se deve

desenvolver um dicionário eletrônico que determine suas paráfrases específicas. A tradução deve ser vista como um uso específico.

Outra utilidade para este modelo é uma espécie de parafraseador automático. Podemos imaginar esse parafraseador como um aperfeiçoamento do recurso que muitos editores de texto possuem de sugerir sinônimos para as palavras destacadas pelo usuário. No parafraseador automático, poderiam ser sugeridas paráfrases para trechos destacados pelo usuário.

Como o tratamento semântico é fundamental para um bom *parser* sintático, esse modelo poderia servir a programas que precisam de uma análise sintática precisa, como parece ser o caso dos revisores gramaticais.

Outro uso do modelo é usá-lo para confeccionar dicionários para rodar em meio digital. Num possível dicionário, o usuário poderia não apenas digitar a palavra cujo significado quer saber ou lembrar ou conferir, como já faz, mas também o usuário poderia ditar o significado cuja palavra correspondente quer saber ou lembrar ou conferir. Ainda poderiam ser feitos dicionários de idéias afins, *thesaurus*, dicionários de sinônimos, ou dicionários de regência a partir de um dicionário eletrônico que seguisse esse modelo.

5
Conclusão

Apesar de incompleto e embrionário, o presente modelo de representação lingüístico tem o mérito de ser uma solução viável para o tratamento semântico com a finalidade de permitir a desejada busca orientada a idéia.

As principais soluções apresentadas pelo modelo são o tratamento do significado como paráfrase e o esquema gráfico criado para formalizar a semântica em acordo com a sintaxe.

Definir o significado de uma palavra ou expressão simplesmente conhecendo as palavras e expressões que as podem substituir simplifica o processo, além de permitir sua automatização. Além disso, a equação entre significado e paráfrase unifica dois problemas do paradigma de busca orientado a palavra. Um problema era que, sem saber quais os significados as palavras requisitadas pelo usuário e os significados das palavras dos textos buscados exerciam, o buscador retornava muitos textos indesejados por apresentar as mesmas palavras requeridas, mas com sentidos diversos aos pretendidos. O segundo problema era que, sem reconhecer as diferentes formas em que uma idéia poderia ser expressa por palavras, o buscador não retornava textos relevantes simplesmente porque não apresentava as mesmas palavras escritas no requerimento do usuário.

A formalização da semântica, tendo em vista as paráfrases, foi desenvolvida a partir de um modelo desenhado exclusivamente para ser usado na busca orientada a idéia. Sendo assim, não só é usado o recurso aos conceitos para representar os significados, mas também foi determinada uma única relação que entre esses conceitos se estabelece. A relação de especificação entre conceitos foi determinada porque se aproxima da relação argumental existente entre as palavras. Desse modo, consegue-se estabelecer uma mesma representação para expressões que também reescrevem palavras.

Por fim, utilizar o grafo para representar as relações entre os conceitos se mostrou hábil por permitir grande flexibilidade, uma vez que no grafo os conceitos não se fixam à ordem, intercalação ou distância entre eles. O que importa é a presença dos conceitos e das relações entre os conceitos. Por isso se pôde tratar os casos em

que uma paráfrase de uma idéia ocorria quando aparecia decomposta nas idéias de várias outras palavras, sendo que essas idéias poderiam ter as mais variadas distribuições entre essas palavras, sendo ainda que essas palavras poderiam estar espalhadas ao longo das diferentes configurações gramaticais.

Além disso, a habilidade de reconhecer grafos como subgrafos de outros grafos permite tratar palavras de mesmo campo semântico que não sejam exatamente paráfrases como as palavras que se relacionam por hiperonímia e hiponímia.

Outras inovações presentes no modelo são importantes por serem resultados encontrados para satisfazer os propósitos a que se servirá o modelo. Entender função sintática como relação argumentativa e a redefinição das classes de palavras são exemplos dessas novidades.

Por esses motivos, o aperfeiçoamento desse modelo se mostra viável para a criação de um buscador orientado a idéia.

Bibliografia

AKHTAR, Shazia / REILLY, Ronan G. / DUNNION, John. *Automating XML markup of text documents*. RECURSO ELETRÔNICO: http://acl.ldc.upenn.edu/N/N03/N03-2001.pdf

AMARAL, Helena Maria Barbosa do. *Estudo de um modelo de rede semântica para banco de dados*. DISSERTAÇÃO DE MESTRADO: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 1978.

ARARIBÓIA, G. *Inteligencia artificial: um curso prático*. Rio de Janeiro : Livros Técnicos e Científicos, 1988.

ASSIS, Patrícia Seefelder de. *Indexação automática por semântica imprecisa*. DISSERTAÇÃO: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 1992.

BORBA, Francisco S. *Uma gramática de valências para o Português*. São Paulo: Editora Ática, 1996.

BOUFADEN, Narjès. *An Ontology-based Semantic Tagger for IE system*. RECURSO ELETRÔNICO: http://acl.ldc.upenn.edu/P/P03/P03-2002.pdf

CAÑAS, Alberto J./ CARVALHO, Marco M. *Mapas Conceituais e IA: uma união improvável*. RECURSO ELETRÔNICO: www.ihmc.us/users/acanas/Publications/ ConceptMapsAI/CanasCarvalhoRBIE2005.pdf

CHOMSKY, Noam. *Aspects of the Theory of Syntax*. Cambridge: The MIT Press, 1965.

COPESTAKE, Ann/ LASCARIDES, Alex/ FLICKINGER, Dan. *An Algebra for Semantic Construction in Constraint-based Grammars*. RECURSO ELETRÔNICO: http://www.cl.cam.ac.uk/~aac10/papers/acl2001.pdf

DIAS, Maria Carmelita Pádua. *Uma proposta de tratamento automático das locuções prepositivas no português*. DISSERTAÇÃO: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 1984.

FELICÍSSIMO, Carolina Howard/ SILVA, Lyrene Fernandes da. BREITMAN/ Karin Koogan/ LEITE, Julio Cesar Sampaio do Pardo. *Geração de Ontologias subsidiada pela Engenharia de Requisitos*. RECURSO ELETRÔNICO: wer.inf.puc-rio.br/WERpapers/artigos/artigos_WER03/carolina_felicissimo.pdf

FERNEDA, Edberto. *Recuperação de informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. TESE: São Paulo: Universidade do Estado de São Paulo, 2003.

FERREIRA, Francisco Eduardo dos Reis. *Desenvolvimento de aplicações baseadas em serviços na web semântica*. DISSERTAÇÃO: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2003.

GREGHI, Juliana Galvani. *Projeto e Desenvolvimento de uma Base de Dados Lexicais do Português*. TESE: Universidade de São Paulo. São Carlos: 2002.

JING, Hongyan. *Usage of WordNet in Natural Language Generation*. RECURSO ELETRÔNICO: http://acl.ldc.upenn.edu/W/W98/W98-0718.pdf

LOH, Stanley. *Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos*. TESE: Porto Alegre: Universidade Federal do Rio Grande do Sul, 2001.

MACKINTOSH, Jennifer. *The Kintsch and Van Dijk model of discourse comprehension and production applied to the interpretation process*. RECURSO ELETRÔNICO: http://www.erudit.org/revue/meta/1985/v30/n1/003530ar.pdf

MOURA, Sabrina Silva de. *Desenvolvimento de interfaces governadas por ontologias para aplicações na Web Semântica*. DISSERTAÇÃO DE MESTRADO: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2004

NEIVA, Sheila Ferreira. *Um modelo de ontologia para verbos do português.* DISSERTAÇÃO: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2000.

OLIVEIRA JR, Osvaldo Novais de Oliveira/ MARTINS, Ronaldo Teixeira/ RINO, Lucia Helena Machado/ NUNES, Maria das Graças Volpe. *O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil*. Série de Relatórios Técnicos do NILC, NILC-TR-01-3, Julho. São Carlos: 2001.

PIETROFORTE, A. Vicente/ LOPES, Ivã. "Semântica Lexical". In: FIORIN, J.L. (Org.). *Introdução à Lingüística*. Vol. II. Princípios de Análise. São Paulo: Ed. Contexto, 2003.

PINTO, Ivone Isidoro. *Uma proposta para recuperação da informação através de redes lexicais:* uma estratégia léxico-quantitativa. TESE: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2002.

PUSTEJOVSKY, James. *The generative lexicon*. Cambridge: MIT Press, 1995.

SANTOS, Maria Angela Moscalewski Roveredo dos. *Extraindo regras de associação a partir de textos*. DISSERTAÇÃO DE MESTRADO: Curitiba: Pontifícia Universidade Católica do Paraná, 2002.

SCHANK, Roger C./ TESLER, Larry. *A conceptual dependency parser for natural language*. RECURSO ELETRÔNICO: http://acl.ldc.upenn.edu/C/C69/C69-0201.pdf

SPECIA, Lucia/ RINO, Lucia Helena Machado. *ConPor: um gerador de estruturas conceituais UNL*. Série de Relatórios Técnicos do NILC, NILC-TR-02-15, Novembro. São Carlos: 2002, 40 p.

SPECIA, Lucia/ RINO, Lucia Helena Machado. *O desenvolvimento de um léxico para a geração de estruturas conceituais UNL*. Série de Relatórios Técnicos do NILC, NILC-TR-02-14, Setembro. São Carlos: 2002, 25 p.

SPECIA, Lucia/ RINO, Lucia Helena Machado. *Representação Semântica: Alguns Modelos Ilustrativos.* Série de Relatórios Técnicos do NILC, NILC-TR-02-12, Julho. São Carlos: 2002, 29p.

SPECIA, Lucia/ RINO, Lucia Helena Machado. *Um gerador de estruturas conceituais UNL para o português.* RECURSO ELETRÔNICO: http://www.nilc.icmc.usp.br/ nilc/ download/Scientia-SpeciaRino.pdf

SUZUKI, Jun/ SASAKI, Yutaka/ MAEDA, Eisaku. *Kernels for Structured Natural Language Data*. RECURSO ELETRÔNICO: http://books.nips.cc/ papers/files/nips16/NIPS2003_AP10.pdf

SZUNDY, Guilherme de Araújo. *Modelagem e implementação de aplicações hipermídia governadas por ontologias para a web semântica*. DISSERTAÇÃO DE MESTRADO: Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro,2004.

VILELA, Mario. *Gramática de valências: teoria e aplicação*. Coimbra: Livraria Almedina, 1992.

Anexo – Corpus

Clássico Fla-Flu termina empatado em 2 a 2

O Flamengo conseguiu um empate por 2 a 2 no clássico realizado em Volta Redonda.

Petkovic deu o passe para o segundo gol, mas perdeu a bola que originou o gol de empate do Flamengo no final do jogo.

Em uma falta de Obina em Leandro, Petkovic cobrou e marcou.

O Flamengo tentava o empate.

O Fluminense teve a chance do segundo gol com Rodrigo Tiuí.

Romeu derrubou Jônatas e o juiz marcou pênalti. Renato bateu no canto esquerdo e empatou.

Petkovic acertou novo chute rente à trave de Diego. Mas enquanto o tricolor vivia de chutes do sérvio, o Flamengo buscava a virada.

Kléber desviou a bola de dois jogadores rubros-negros.

O Flamengo levou o segundo gol. Petkovic fez belo passe para Tuta, que invadiu a área e bateu forte. Diego ainda tocou na bola, que morreu na rede.

Mas o Flamengo não desistiu e empatou. Diego Souza bateu de fora da área. A bola passou por debaixo de Kléber e entrou.

O Flu teve a chance de desempatar com Leandro.

Em clássico emocionante, Fla-Flu acaba empatado

O clássico carioca aconteceu em Volta Redonda.

O time rubro-negro deu espaços ao time tricolor. Tuta bateu na saída de Diego, abrindo o placar.

A equipe tricolor se acomodou com a vantagem.

Rodrigo Tiuí chutou para fora. O clube da Gávea quase empatou.

André saiu jogando errado e deu a bola nos pés de Petkovic. O sérvio chutou rente à trave de Diego.

A equipe flamenguista empatou. Jônatas arrancou e foi derrubado por Romeu dentro da área. Na cobrança, Renato marcou seu sexto gol no campeonato.

A igualdade acordou o Fluminense.

Os tricolores marcaram o segundo gol. Petkovic achou Tuta na entrada da área e o centroavante não desperdiçou.

Diego Souza chutou de longe e Kleber aceitou.

O Fluminense quase marcou o terceiro, mas o chute de Leandro esbarrou na trave.

Flamengo e Fluminense empatam em Volta Redonda

Flamengo e Fluminense empataram por 2 a 2, em Volta Redonda (RJ), em clássico carioca válido pelo Campeonato Brasileiro.

Tuta abriu o placar para o Fluminense. O Flamengo empatou com Renato. Tuta desempatou a partida. Mas o Flamengo chegou ao empate, com Diego Souza.

O Fluminense volta a jogar pelo Campeonato Brasileiro, quando enfrenta o Santos, em Volta Redonda (RJ). O Flamengo encara o Corinthians, no Rio.

Fla e Flu empatam em jogo emocionante

Tuta marcou os dois gols do tricolor, enquanto Renato e Diego Souza descontaram para o rubro-negro.

As duas equipes voltam a campo neste domingo. O Flamengo recebe o Timão no Luso-Brasileiro, enquanto o tricolor enfrenta o Santos, em Volta Redonda.

O Fluminense foi logo abrindo o placar. Após uma cobrança de falta de Petkovic, a bola bateu na barreira e voltou para Preto Casagrande. O apoiador lançou Tuta, que chutou na saída de Diego.

Os jogadores do Flamengo buscavam o gol de empate.

Com a vantagem no placar, o tricolor preparava melhor os lances de ataque.

O jogo melhorou muito, e o Fluminense quase ampliou em chute de Petkovic.

O empate do Flamengo aconteceu. Na entrada da área, Jônatas foi derrubado e o juiz marcou pênalti. Renato cobrou e deslocou o goleiro Kléber.

O tricolor ampliou. Petkovic deu passe perfeito para Tuta, que invadiu a área e chutou forte.

Apesar da vantagem tricolor no placar, Diego Souza chutou forte de fora da área e a bola passou por baixo do goleiro Kléber, deixando tudo igual no placar.

As duas equipes seguiram buscando a vitória até o fim.

Fla e Flu empatam em clássico emocionante: 2 a 2

Gol do ex-tricolor Diego Souza evita derrota rubro-negra

Flamengo e Fluminense fizeram no Brasileiro um clássico emocionante, com gols. O Tricolor esteve mais próximo da vitória, mas o Rubro-Negro empatou o jogo por 2 a 2. Tuta fez os gols do time das Laranjeiras. Renato e Diego Souza marcaram para a equipe da Gávea.

O clássico foi em Volta Redonda.

O Tricolor aproveitou a bola parada para abrir o placar. Pet colocou no canto de Diego. Tuta dominou para bater cruzado e fazer 1 a 0.

O Rubro-Negro se fartou de perder gol por não ter um centroavante bem posicionado para concluir as jogadas.

A zaga rubro-negra errou feio e Pet quase ampliou. Mas o time da Gávea tentava empatar a partida. E conseguiu por meio de uma jogada polêmica: Jônatas foi derrubado por Romeu e o árbitro marcou pênalti. Renato bateu bem e empatou a partida: 1 a 1.

Os lances perigosos eram despediçados pelos dois lados. Quando parecia que o Flamengo iria virar a partida, Tuta recebeu belo passe de Pet e bateu forte: 2 a 1.

O time da Gávea conseguiu empatar com um atleta bem conhecido nas Laranjeiras. Diego Souza - revelado pelo Tricolor - arriscou de fora da área e Kléber levou um frango.

O clássico ainda teve uma bola na trave de Pet, mas o resultado acabou com um empate emocionante em Volta Redonda.

Emoção até o final

O empate em 2 a 2 não foi um bom resultado para o Fluminense, em Volta Redonda. Tuta (dois), Renato e Diego Souza marcaram os gols.

Num jogo de alta tensão, saiu o primeiro gol depois que Petkovic bateu falta na barreira, a bola voltou para Preto, que centrou para Tuta dominar no peito e chutar na saída de Diego.

Apesar da desvantagem no placar, o Flamengo insistiu.

A pressão rubro-negra continuou. A arbitragem marcou pênalti de Romeu em Jônatas, convertido por Renato.

A partida melhorou. Flamengo e Fluminense lutaram o quanto puderam, mas apenas empataram, com gols de Tuta (o atacante é o artilheiro do Flu, com 13 gols), e Diego Souza, numa falha grotesca de Kléber.

Flamengo arranca empate contra o Fluminense

Em uma partida recheada de emoção, Fluminense e Flamengo empataram em 2 a 2, em Volta Redonda. Tuta marcou os dois gols do Fluminense, que joga com o Santos no domingo. Já Renato e Diego Souza anotaram para o Flamengo, que enfrenta o Corinthians.

Petkovic chegou a marcar em cobrança de falta. Na repetição da cobrança, o meia bateu mal, mas no rebote Tuta foi lançado por Preto Casagrande e chutou cruzado para colocar o Fluminense em vantagem no clássico.

O gol colocou ainda mais emoção na partida. Léo Mattos quase empatou para o Flamengo. O Tricolor respondeu, mas desperdiçou boa chance.

O Rubro-Negro buscava o empate a todo momento.

Petkovic quase marcou. O Flamengo foi beneficiado pela arbitragem, que marcou pênalti quando Jônatas foi derrubado fora da área. Na cobrança, Renato empatou a partida.

O Fluminense recorreu ao talento de Petkovic. O meia colocou Tuta livre na área e o atacante não desperdiçou a oportunidade de colocar o Tricolor novamente em vantagem no placar.

Diego Souza arriscou de fora da área e Kléber falhou. 2 a 2 no clássico.

Leandro ainda mandou uma bola na trave, mas não conseguiu impedir que o empate prevalecesse em Volta Redonda.

Empate num Fla-Flu de gols, expulsões e emoção

Rivais protagonizam um clássico eletrizante e empataram por 2 x 2 em Volta Redonda

As duas equipes protagonizaram um clássico sensacional, em Volta Redonda. Tuta fez os dois gols tricolores, enquanto Renato, de pênalti, e Diego Souza marcaram para o time rubro-negro.

O Flu recebe o Santos, em Volta Redonda, enquanto o Fla enfrenta o Corinthians, no Rio.

Após uma cobrança de falta, Tuta matou no peito e chutou cruzado, abrindo o placar.

Inferiorizado no placar, o Flamengo partiu para o ataque.

Jônatas foi derrubado por Romeu fora da área, mas o árbitro marcou pênalti. Renato cobrou bem e empatou a partida.

Petkovic perdeu a chance de fazer o segundo gol tricolor.

Animado pelo gol de empate, o Flamengo começou a pressionar em busca do segundo.

Foi o Fluminense que chegou ao segundo gol, quando Petkovic deu passe na medida para Tuta chutar forte, sem chance para Diego, que ainda tocou na bola, mas não evitou o gol.

Mas o Flamengo não desistiu e chegou ao empate, com um *frango* de Kleber num chute de Leandro. Leandro ainda acertou a trave, mas o empate foi o resultado de um Fla-Flu emocionante.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo