

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 textos¹, um arquivo de texto teria também as informações sobre os significados depreendidos no texto. É sobre essas informações semânticas² 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 palavras³.

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, todas 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 idéia⁴.

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 dele⁵.

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 apreendido 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 homônimas e não polissemias⁶.

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

João	matou	José
A	B	C

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 homônimas e não polissemias, 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.

José	matou	João
C	B	A

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 presente.

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.

O assassinato de José por João		
B	C	A

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 referencialidade⁷.

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 nomes⁸.

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 entidades⁹ 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 evento¹⁰.

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 contidos¹¹.

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 completar¹². 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 distinta¹³.

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 consequência, 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 compõem.

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

João	causou	a morte	de	José
A	B	C	D	

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:

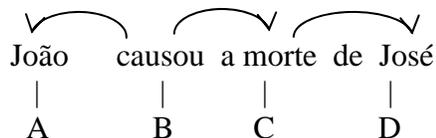


Figura 5 – relação argumental

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 idéias 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:

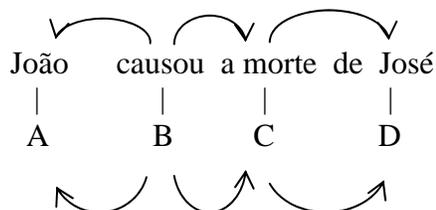


Figura 6 – significados

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”.

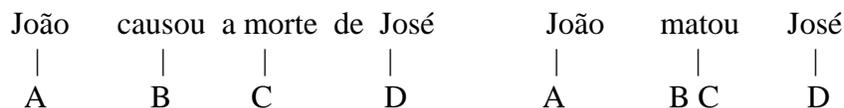


Figura 7 – redistribuindo os significados

Se formos representar as relações argumentais de “João matou José” por setas, e por analogia, atribuímos 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 aos significados das palavras de ambas as frases devem ser idênticas. Sendo assim, teremos:

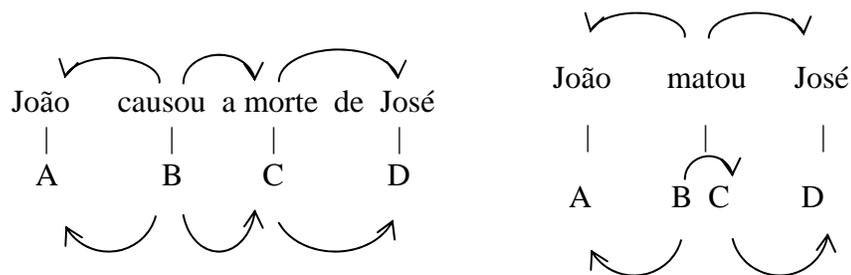


Figura 8 – atribuindo setas iguais às paráfrases

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.

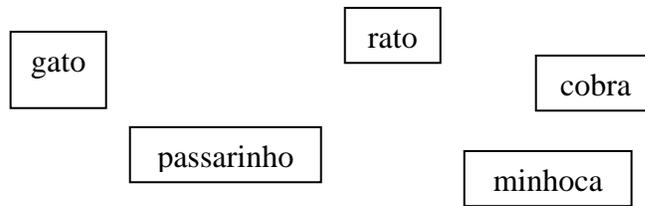


Figura 9 – nomes de animais

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:

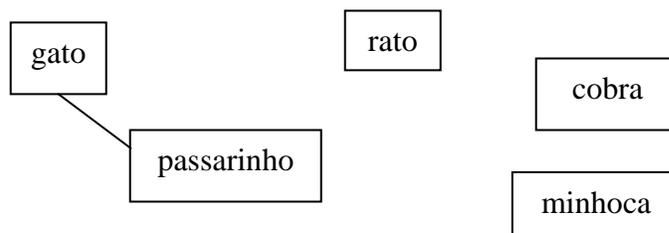


Figura 10 – a relação caça

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

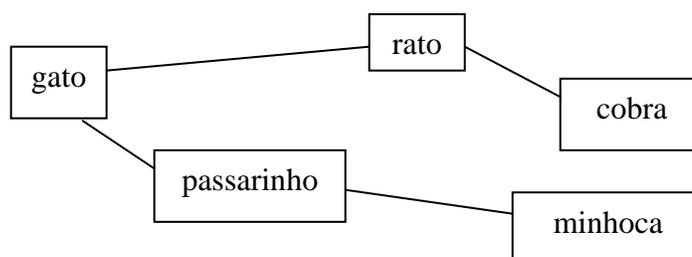


Figura 11 – grafo da caça

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.

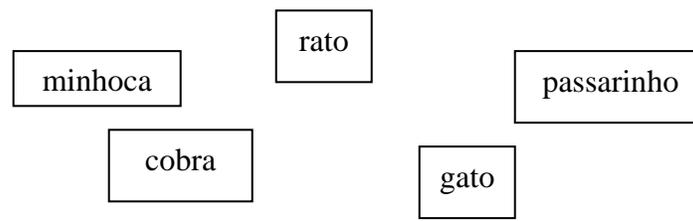


Figura 12 – grafo com outra configuração

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:

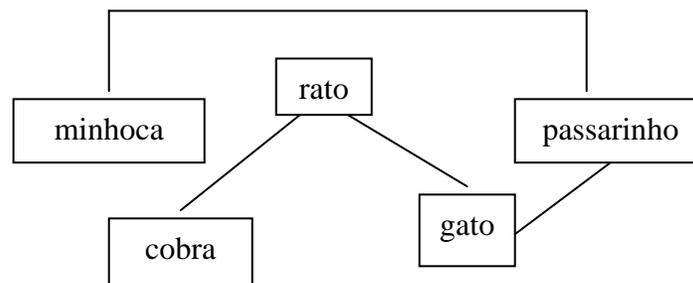


Figura 13 – novo grafo de caça

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.

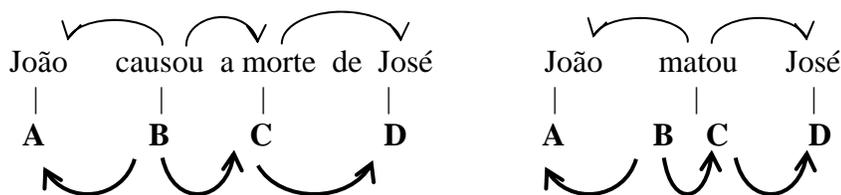


Figura 14 – grafos congruentes

Nosso grafo é um tipo particular de grafo, em que o arco (a relação) é orientado, isso é, é representado por uma seta¹⁴. 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 mudam¹⁵. Se as relações mudam, mudam também as representações. Vejamos:

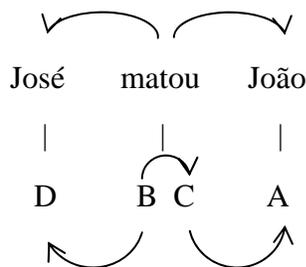


Figura 16 – grafo de uma sentença não paráfrase

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 homônimas.

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:

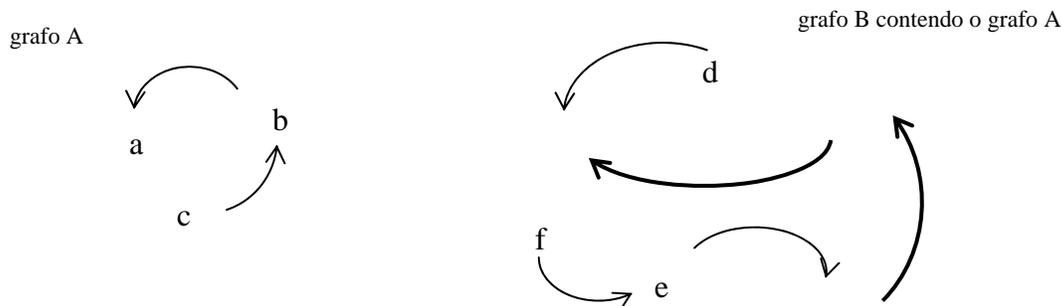


Figura 17 – um grafo contendo outro grafo

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:

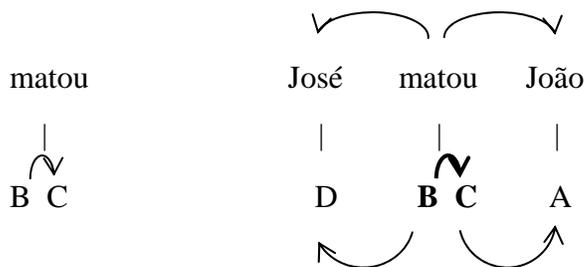


Figura 18 – subgrafo

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:

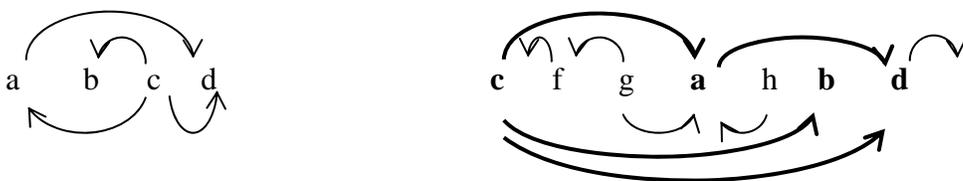


Figura 19 – habilidades de um grafo

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” conterà 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:

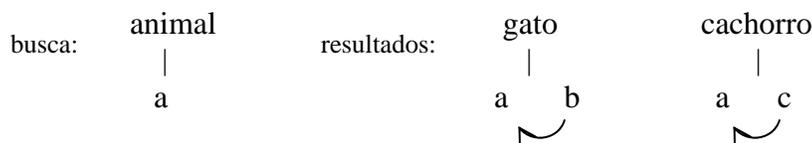


Figura 20 – grafo de hipônimos contendo o grafo de hiperônimos como subgrafo

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 concluimos 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 a *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:

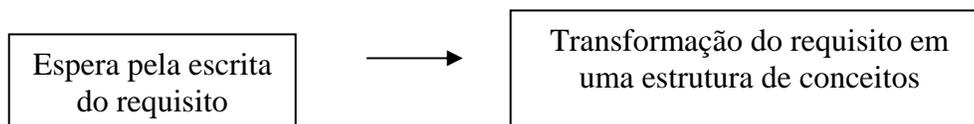


Figura 22 – algoritmo 2

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.

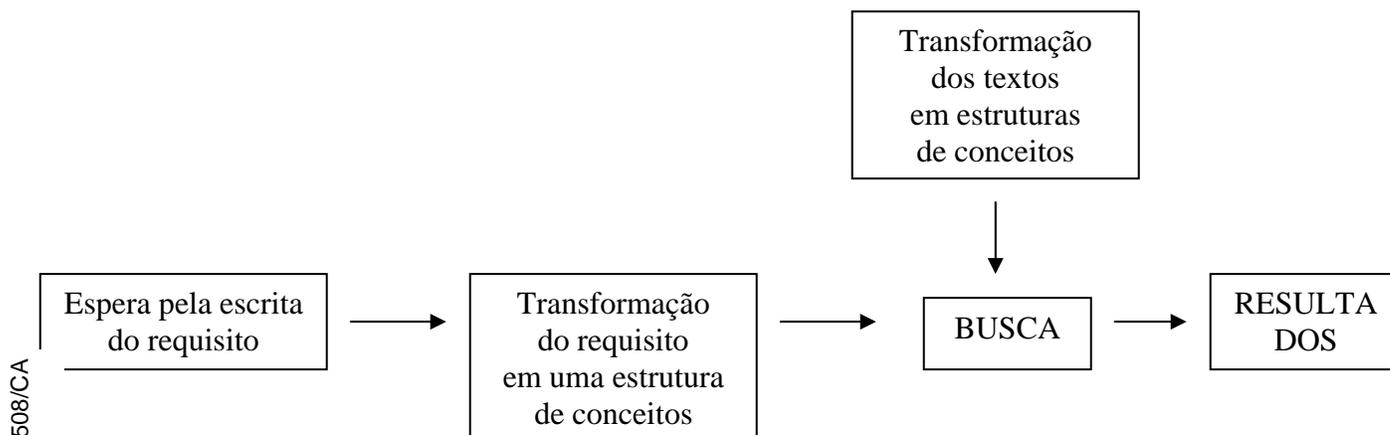


Figura 23 – algoritmo 3

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.

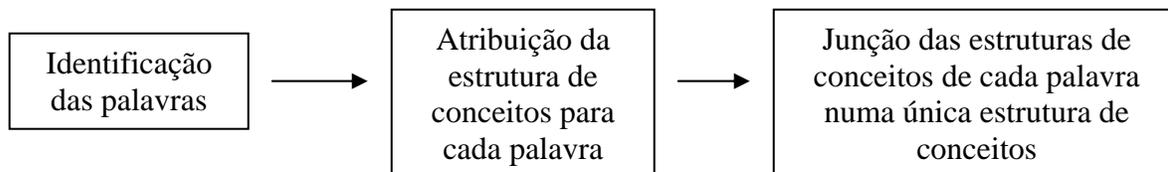


Figura 24 – algoritmo 4

Para atribuímos 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 eletrônico¹⁶ 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.

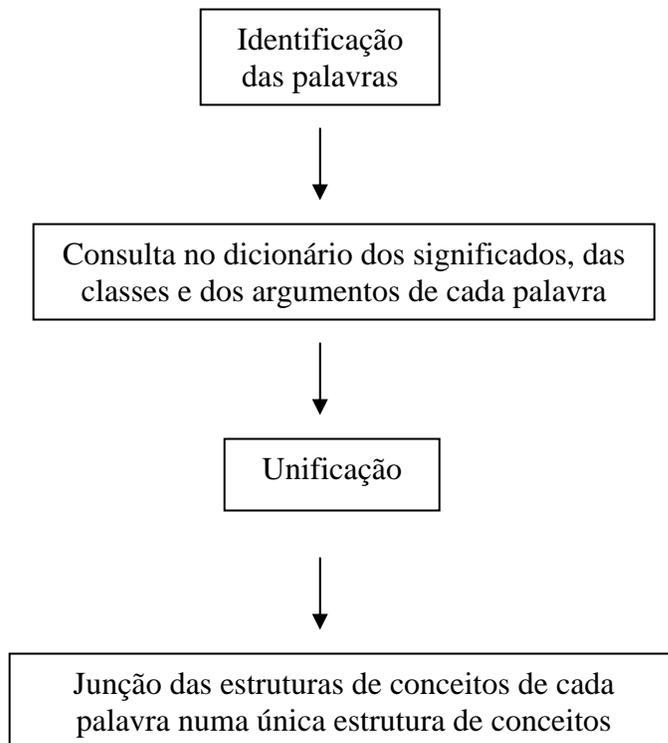


Figura 25 – algoritmo 5

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 visto¹⁷.

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 indireta¹⁸. 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 atribuímos 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 sintagma¹⁹.

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.

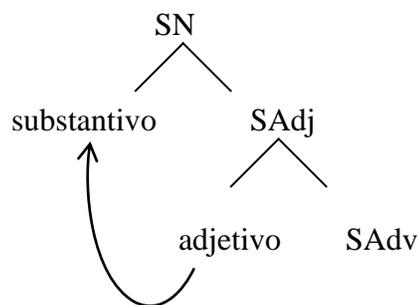


Figura 26 – núcleos do sintagma adjetivo e nominal

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-barr 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.

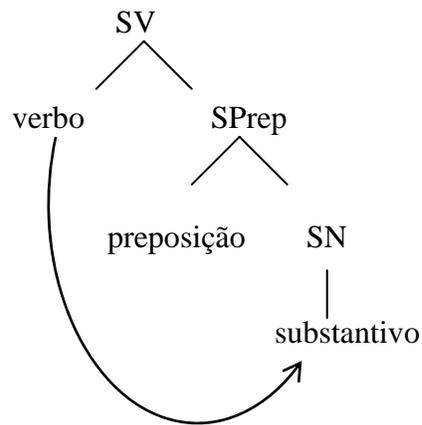


Figura 27 – ignorando o núcleo do sintagma preposicional

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.

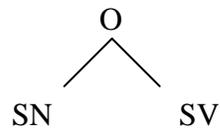


Figura 28 – hierarquia entre sintagma nominal do sujeito e sintagma verbal

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.

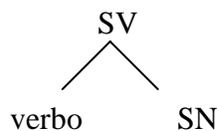


Figura 29 – hierarquia entre sintagma nominal do objeto e sintagma verbal

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.

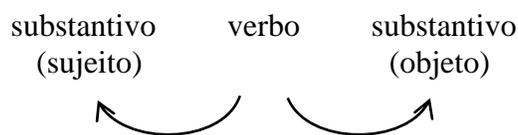


Figura 30 – hierarquia entre os argumentos do verbo

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.

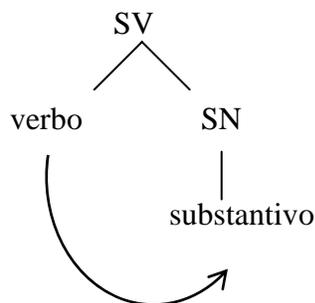


Figura 31 – hierarquia entre verbo e seu argumento

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.

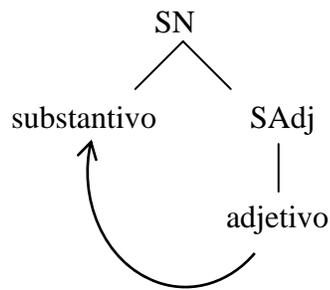


Figura 32 – hierarquia entre o adjetivo e seu argumento

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”).

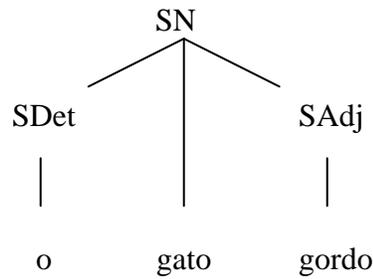


Figura 33 – estrutura do sintagma substantivo

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

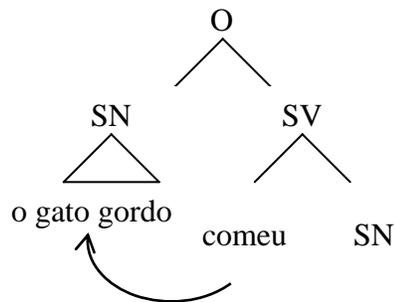


Figura 34 – sintagma nominal como argumento do verbo

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 animado²¹. 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 argumental²³. 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.

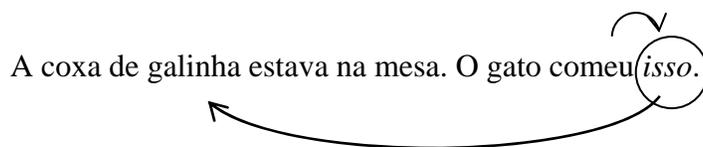


Figura 36

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.

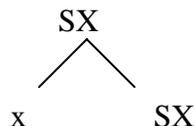


Figura 37 – estrutura de um sintagma com recursão

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”.

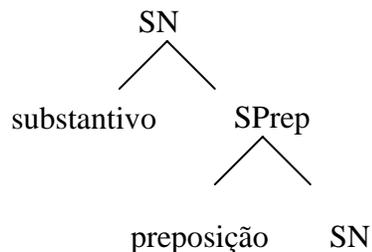


Figura 38 – recursão no sintagma nominal

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.

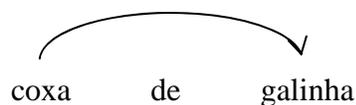


Figura 39 – sintagma preposicionado como argumento do substantivo

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.

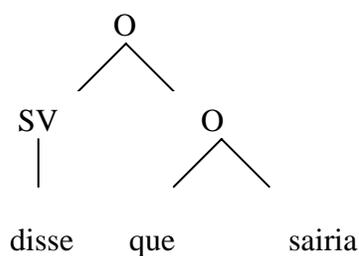


Figura 40 – recursão no sintagma verbal

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.

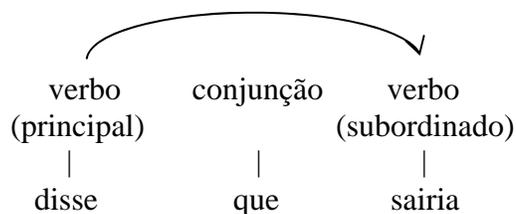


Figura 41

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 exigências²⁵ 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 estabeleçam 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 argumentos²⁶. 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.

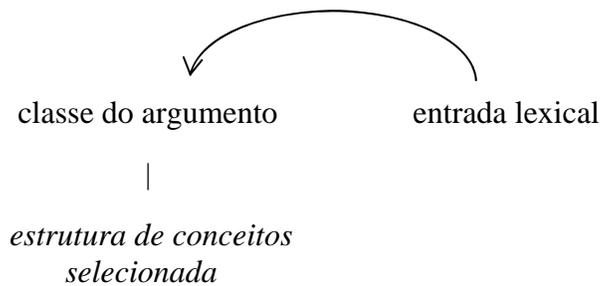


Figura 42 – representação da estrutura de argumentos

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:

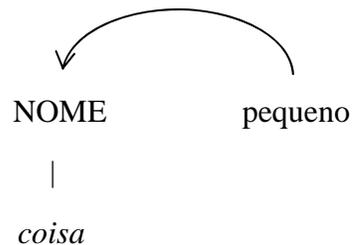


Figura 43 – exemplo de estrutura de argumento

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:

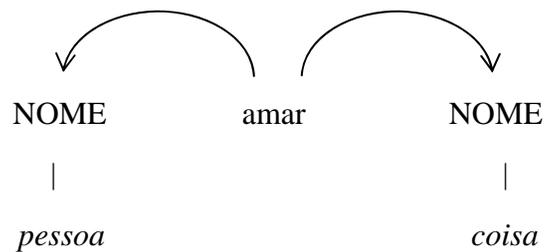


Figura 44 – seleção semântica

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:

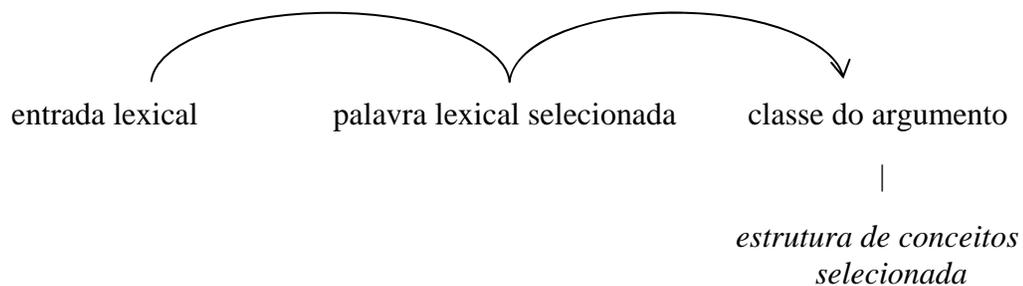


Figura 45 – representação da seleção semântica na estrutura de argumentos

Eis um exemplo disso:

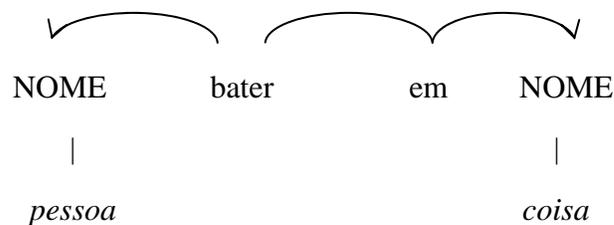


Figura 46 – exemplo de seleção semântica

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:

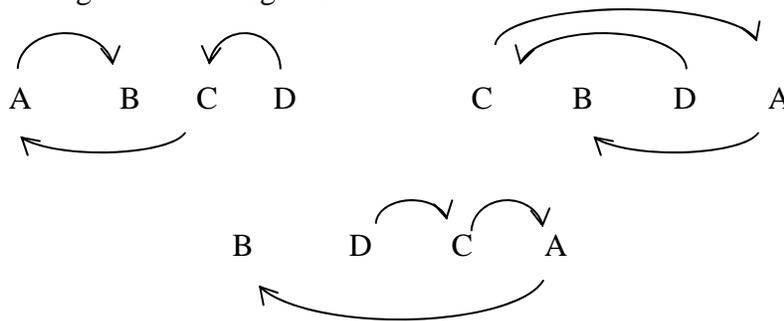


Figura 47 – grafos congruentes

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:

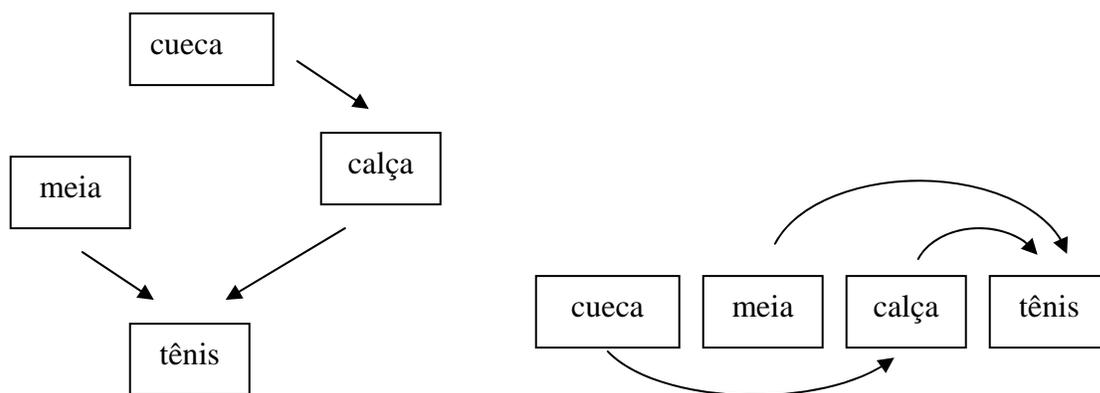


Figura 48 – ordenação topológica

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 nosso 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:

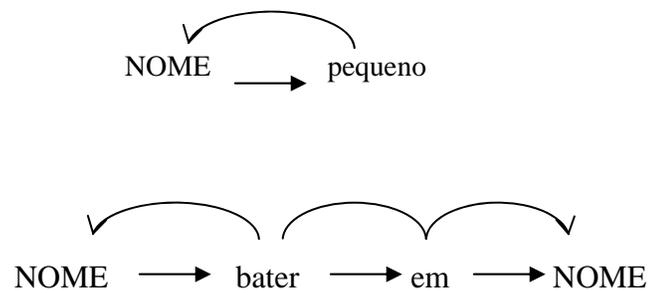


Figura 49 – ordenação topológica na estrutura de argumentos

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 parafraseiem²⁷. 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 sentido 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 palavras²⁸. 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 tranquilamente.

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

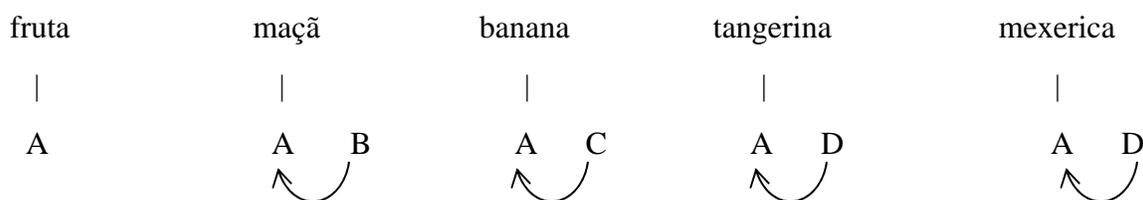


Figura 51 – exemplo de atribuição de conceitos

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 terei a escolher formas mais mnemônicas para designar um conceito.

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

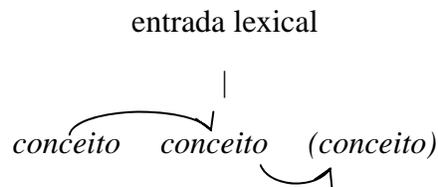


Figura 52 – representação da estrutura de conceitos

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.

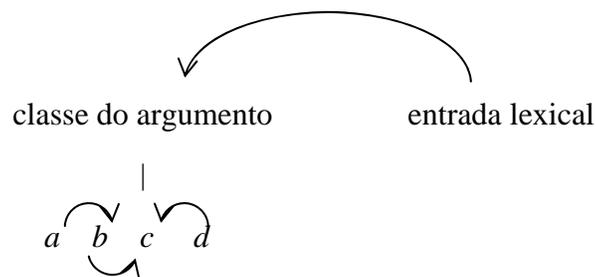


Figura 53 – estrutura de argumentos selecionando um grafo

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 argumentos³⁰. 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”.

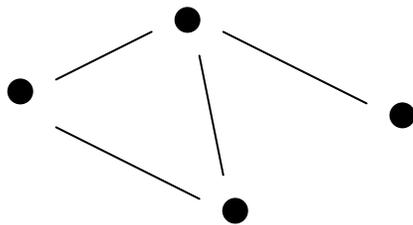


Figura 54 – exemplo de grafo conexo

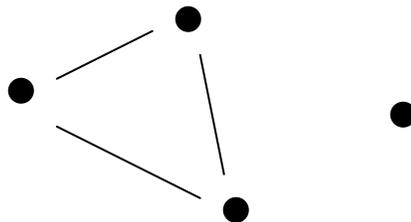


Figura 55 – exemplo de grafo desconexo

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”.

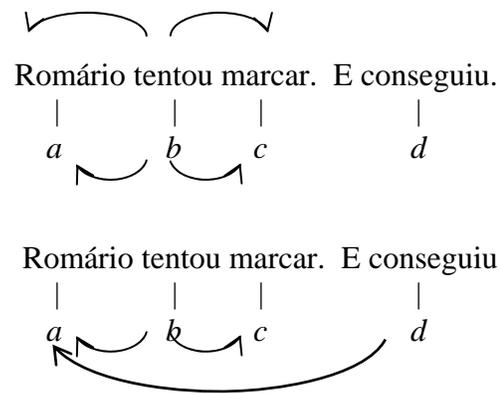


Figura 56 – coordenação 1

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”.

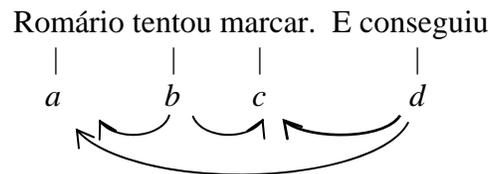


Figura 57 – coordenação 2

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

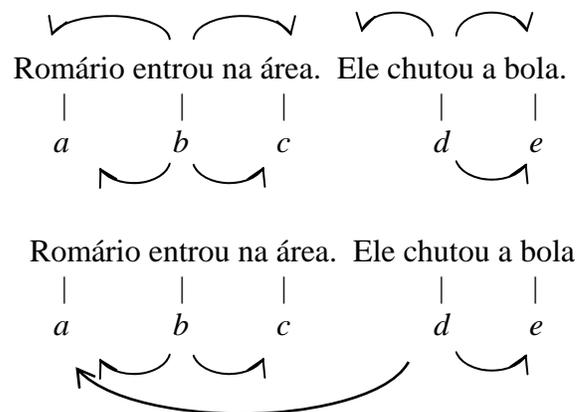


Figura 58 – anáfora 1

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

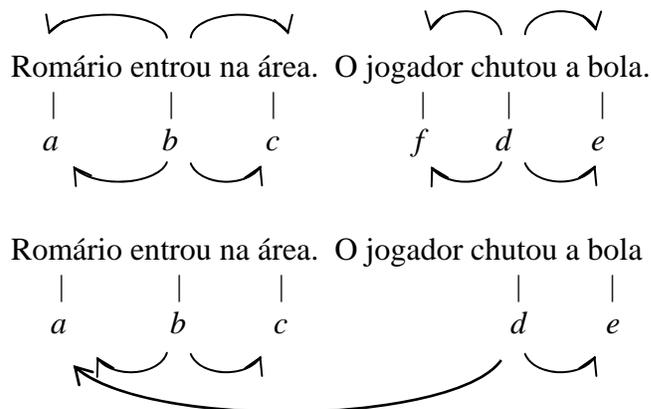


Figura 59 – anáfora 2

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.