

T-Lex: Thesaurus com Estruturação Semântica e Operações Gerativas

Marco Gonzalez

PUCRS - Faculdade de Informática
Av.Ipiranga, 6681 – Prédio 16, térreo
90619-900 Porto Alegre, Brazil
gonzalez@inf.pucrs.br

Vera Lúcia Strube de Lima

PUCRS - Faculdade de Informática
Av.Ipiranga, 6681 – Prédio 16, térreo
90619-900 Porto Alegre, Brazil
vera@inf.pucrs.br

ABSTRACT

T-Lex is a thesaurus manually constructed from a traditional dictionary, adopting aspects from the Generative Lexicon Theory (GLT) proposed by James Pustejovsky, as well as object-oriented software modeling concepts. The Qualia structure of GLT is adapted for describing nouns, verbs and adjectives on a differentiated way, respectively, as objects, operations, and states. Generative operations which return sets of related words are included in the T-Lex prototype implementation. So, this thesaurus is a helpful tool in of automatic query expansion for information retrieval in Portuguese language.

Keywords: natural language processing, artificial intelligence, thesaurus, data mining, generative operations

RESUMO

T-Lex é um thesaurus construído manualmente a partir de dicionário tradicional, adotando aspectos da Teoria do Léxico Gerativo (TLG) de James Pustejovsky e conceitos da modelagem de software orientada a objetos. A estrutura Qualia da TLG é adaptada com o objetivo de descrever substantivos, verbos e adjetivos de forma diferenciada, respectivamente, como objetos, operações e estados. Operações gerativas que visam a geração de campos lexicais são incluídas na implementação do protótipo do T-Lex e, dessa forma, o thesaurus é utilizado com bons resultados na expansão automática de consulta em sistema de recuperação de informação em língua portuguesa.

Palavras-chave: processamento de linguagem natural, inteligência artificial, thesaurus, base de dados, operações gerativas

1 INTRODUÇÃO

No campo da informação e da documentação, os thesauri têm absorvido diversas funções [15][11], entre as quais são citadas as de:

- seleção de vocabulário, durante a produção de textos,
- normalização de termos para representação de conceitos, durante a indexação de documentos,
- caracterização de temas, durante *clustering* de documentos, e
- associação de termos, durante a realimentação de consulta em recuperação de informação (RI).

Especificamente, em RI, encontram-se aplicações de thesauri durante todo o período da história de desenvolvimento desses sistemas [4][14][7]. Também, no âmbito das bibliotecas tradicionais, em tarefas de catalogação e pesquisa bibliográfica, o thesaurus vem sendo discutido como instrumento fundamental para controle de vocabulário [16][6].

O thesaurus que apresentamos aqui, o T-Lex, possui uma estruturação semântica projetada para implementar relacionamentos lexicais, levando em conta aspectos da Teoria do Léxico Gerativo (TLG) de Pustejovsky [9] e conceitos de modelagem de software orientada a objetos (MSOO) [12]. Partindo desses fundamentos, é implementado um thesaurus tornando explícitas as relações semânticas, nas representações lexicais. Para tanto, são introduzidas referências diretas (na forma de descritores) aos itens lexicais relacionados ao termo descrito. É dado tratamento diferenciado para as classes gramaticais dos substantivos, dos verbos e dos adjetivos. Operações gerativas específicas (especialização, co-herança, associação, equivalência, decomposição e agregação) são idealizadas e implementadas, viabilizando o poder de geração de campo lexical do T-Lex.

Na seção 2 é feita uma breve referência à teoria do léxico gerativo com ênfase na estrutura Qualia; na seção 3 são apresentadas, quanto a esta estrutura, as adaptações propostas neste trabalho e adotadas no thesaurus desenvolvido, sendo também abordadas diferenças em relação à estrutura original; na seção 4 é discutida a implementação do T-Lex quanto à linguagem de descrição e à codificação dos itens lexicais; na seção 5 são descritas as operações gerativas incluídas no T-Lex como recurso para geração de campo lexical, é analisada a composição dos conjuntos-resposta produzidos por essas operações, e é apresentada uma aplicação do T-Lex para expansão automática de consulta em sistemas de RI, incluindo resultados preliminares; na seção 6 são tecidas considerações finais sobre o presente trabalho.

2 TEORIA DO LÉXICO GERATIVO

A TLG introduz um conjunto de recursos para análise semântica de expressões em linguagem natural. Um léxico fundamentado nessa teoria, segundo seu autor, é útil na abordagem de questões como: a explicação da natureza polimórfica da linguagem; a caracterização da semanticalidade de expressões em linguagem natural; a captura do uso criativo das palavras em novos contextos; e o desenvolvimento de uma representação semântica co-composicional mais rica. Para maiores detalhes, ver [9].

2.1 Organização do Léxico na TLG

Um léxico semântico, de acordo com a TLG, é caracterizado como um sistema computacional envolvendo quatro níveis de representação: estrutura de Argumentos, estrutura de Eventos, estrutura Qualia e estrutura de Herança Lexical. Um conjunto de operações gerativas, que se agregam ao léxico, conectam esses quatro níveis, possibilitando derivações composicionais das palavras no contexto.

A estrutura de argumentos apresenta informações sobre o comportamento sintático de um item lexical, numa representação gramatical em relação a um predicado. Por exemplo, construir possui como argumentos indivíduo animado, artefato e materiais, enquanto livro tem uma estrutura de argumentos composta por informação e objeto físico.

A estrutura de eventos define os tipos de eventos envolvidos com um item lexical, podendo incluir uma estrutura de subeventos, se houver, além da identificação do evento predominante e da restrição temporal entre eles. Por exemplo, construir apresenta subeventos dos tipos processo e estado, sendo o primeiro anterior, no tempo, e predominante em relação ao segundo.

A estrutura de herança lexical identifica o modo como um item lexical está relacionado a outros, num tipo de malha ou rede. Assim, é determinada a contribuição de cada item lexical para a organização global do léxico. A estrutura de herança lexical é produzida essencialmente pela estrutura Qualia, descrita a seguir.

2.2 Estrutura Qualia na TLG

Neste trabalho, colocamos especial destaque sobre a estrutura Qualia, já que a sua organização em quatro papéis permite, através deles, estabelecer o relacionamento dos outros níveis de representação (Argumentos e Eventos) entre si. Na descrição da Qualia de um item lexical α , na TLG, temos os seguintes papéis:

- Formal: que distingue α num amplo domínio;
- Constitutivo: que descreve o que faz parte de α ;
- Agentivo: que especifica como α passou a existir; e
- Télico: que explica qual a função ou o propósito de α .

O quadro 1 apresenta, como exemplo, a estrutura Qualia do item lexical livro, onde: e_1 e e_2 são variáveis que representam eventos da estrutura de eventos; v e w são variáveis que representam entidades envolvidas com os respectivos eventos; e o símbolo \bullet indica um paradigma conceitual lexical que descreve, no caso, a carga semântica simultânea de objeto-físico e informação, fazendo com que livro herde o comportamento funcional de ambos.

<p><u>Livro</u> Formal: <u>suporte</u> (<u>objeto-físico</u>, <u>informação</u>) Constitutivo: <u>capa</u>, <u>folha</u>, <u>texto</u>, ... Agentivo: <u>escrever</u> (e_2, v, <u>objeto-físico</u>\bullet<u>informação</u>) Télico: <u>ler</u> (e_1, w, <u>objeto-físico</u>\bullet<u>informação</u>)</p>
--

Quadro 1. Qualia do item lexical livro na TLG

3 ESTRUTURA QUALIA ADAPTADA NO T-LEX

Na TLG, a estrutura Qualia tem uma constituição única, independentemente da categoria gramatical do item lexical descrito. O que pode acontecer é que nem todos os papéis sejam preenchidos em todos os casos.

Diferentemente do proposto originalmente na TLG, no T-Lex cada categoria gramatical possui uma Qualia específica. As categorias gramaticais consideradas no T-Lex são: substantivo comum concreto, substantivo abstrato, substantivo próprio, verbo e adjetivo. Nas descrições que seguem, as definições adotadas para as categorias gramaticais são aquelas da gramática corrente para o português [1], e os conceitos de orientação a objetos utilizados baseiam-se na MSOO [12].

Substantivo comum concreto (SCC)

Os substantivos comuns são os que designam seres da mesma espécie, como ferramenta, e os substantivos concretos são os que designam seres de existência real ou que a imaginação apresenta como tais, como computador, e método. Por outro lado, uma classe de objetos, segundo a MSOO, é definida como um grupo de objetos com propriedades (atributos) semelhantes, o mesmo comportamento (operações), as mesmas ligações (relacionamentos com outros objetos) e, por consequência, a mesma semântica. Por analogia, um SCC engloba estas características, representando uma classe de objetos reais ou imaginários. Assim, no T-Lex, os SCCs são descritos por estruturas Qualia que podem ter todos ou alguns dos seguintes papéis:

- **Formal.** Consiste na generalização dos objetos da classe descrita através de outra classe representada, também, por um SCC. Exemplo: representação, como formal de sinal.
- **Constitutivo.** Indica a constituição dos objetos da classe descrita através de outras classes representadas por SCCs e/ou SABs, especificando componentes, partes ou constituintes. Exemplo: sintaxe, como constitutivo de gramática.
- **Agentivo.** Consiste na especificação do que faz existir os objetos da classe descrita. Pode apresentar as seguintes informações:
 - sobre a **criação** dos objetos da classe descrita através de operações executadas por objetos de outras classes. Exemplo: estruturar, como agentivo (criação) de estrutura.
 - sobre a **inicialização** dos objetos da classe descrita através de operações executadas pelos próprios objetos desta classe. Define o que estes objetos fazem ao existirem (ou para passar a existirem). Exemplo: usar, como agentivo (inicialização) de usuário.
- **Télico.** Consiste na caracterização, através de operações, especificando funções ou propósitos dos objetos da classe descrita. Exemplo: processar, como télico de processador.

Substantivo abstrato (SAB)

Os substantivos abstratos designam qualidades ou estados, como normalidade, sentimentos ou sensações, como calor, e atos, como instalação. Tais qualidades ou estados, sentimentos ou sensações e atos são ditos em relação aos seres, dos quais se podem abstrair e sem os quais não poderiam existir [1]. Objetos, segundo a MSOO, não incluem apenas entidades físicas, como empregado, mas também “conceitos”, como pagamento. Os SABs, portanto, representam estas últimas classes de objetos e, no T-Lex, são descritos por estruturas Qualia que podem ter todos ou alguns dos seguintes papéis:

- **Formal.** Consiste na generalização dos objetos da classe descrita através de outra classe representada, também, por um SAB. Exemplo: relação, como formal de sincronismo.
- **Constitutivo.** Indica a constituição dos objetos da classe descrita através de outras classes representadas por SABs e/ou SCCs, especificando componentes, partes ou constituintes. Exemplo: subordinação e nível, como constitutivos de hierarquia.
- **Agentivo.** Identifica o responsável pela existência dos objetos da classe descrita e de onde estes podem ser abstraídos. Especifica objetos reais ou imaginários, através de SCC, a quem (ou a que) se atribui a qualidade, o sentimento ou a sensação, o ato, ou o estado que se descreve. Exemplos: construtor, como agentivo de construção.
- **Télico.** Identifica as ações correspondentes, na forma de operações, que representam funções ou propósitos que se pode associar aos objetos da classe descrita ou aos “seres” relacionados a tais objetos. Exemplo: processar, como télico de processamento.

Substantivo próprio (SPR)

Os substantivos próprios, em oposição aos substantivos comuns, são aplicados cada um a um ser em particular, como Venezuela. Cada classe de objetos, segundo a MSOO, descreve um conjunto possivelmente infinito de objetos individuais e cada um deles é dito ser uma instância de sua classe e, assim, são classificados. Os SPRs são representados, no T-Lex, como instâncias de classes de objetos, e são descritos por estruturas Qualia simples com um único papel:

- **Formal.** Consiste na classificação do objeto descrito através de uma classe representada por um SCC. Exemplo: mês (na descrição de Abril) e cidade (na descrição de Mérida).

Verbo (VRB)

Os verbos exprimem ação, como executar, estado, como estar, ou fato (ou fenômeno), como chover. Uma operação, conforme a MSOO, é uma função, como calcular, ou uma transformação, como atualizar, que pode ser aplicada a (ou por) objetos de uma classe. Assim, representam-se os verbos, aqui, como operações, no sentido de expressar a dinâmica, em termos de funções e propósitos, relacionada a objetos reais, imaginários ou abstratos. No T-Lex, os VRBs são descritos por estruturas Qualia que podem ter todos ou alguns dos seguintes papéis:

- **Formal.** Consiste na generalização da operação descrita através de outra operação representada, também, por um VRB. Exemplo: criar (na descrição de modelar).
- **Constitutivo.** Indica a constituição da operação descrita, descrevendo-a através de outras operações, também VRBs, necessárias para efetivá-la. Exemplo: chegar e ter(acesso), como constitutivo de alcançar.
- **Agentivo.** Especifica a entrada da operação, na forma de argumentos, representados por objetos reais, imaginários ou abstratos, através de SCCs e SABs. Exemplo: inteligência, como agentivo de estudar¹.
- **Télico.** Especifica a saída da operação, na forma de objetos reais, imaginários ou abstratos, através de SCCs e SABs. Exemplo: ordenação e ordem, como télico de ordenar.

Adjetivo (ADJ)

Os adjetivos expressam qualidades ou características dos seres, como amigável. Um estado, de acordo com a MSOO, é uma abstração dos valores dos atributos (estado adolescente, associado ao atributo idade) e das ligações (estado casado, associado ao relacionamento casamento) de um objeto, podendo estar associado a uma atividade (estado verificador, associado à atividade verificar) ou a uma condição (estado feliz, associado à condição felicidade). Um ADJ representa, portanto, um estado que pode ser atribuído aos objetos de uma classe, e no T-Lex os adjetivos são descritos por estruturas Qualia que podem ter todos ou alguns dos seguintes papéis:

¹ Estudar significa “aplicar a inteligência para aprender ...” [2].

- **Formal.** Consiste na generalização do estado descrito através de um “superestado” representado por um outro ADJ. Exemplo: útil, como formal de utilizado.
- **Constitutivo.** Indica a constituição do estado descrito através de outros estados, também ADJs, que podem estar contidos nele ou ter interseção com ele. Exemplo: independente e incondicional, como constitutivo de absoluto.
- **Agentivo.** Consiste na especificação do que faz existir o estado descrito. Em analogia aos três elementos (evento, condição e ação) possíveis numa transição de estados na MSOO [12], pode apresentar as seguintes informações:
 - sobre a **causação** do estado descrito através de operações executadas por objetos de outras classes que não a relacionada a este estado. Exemplo: estruturar, como agentivo (causação) de estruturado.
 - sobre a **condição**, expressa na forma de SABs, que deve ser verdadeira (ou seja, deve existir) para que o estado descrito também possa existir. Exemplo: processamento, como agentivo (condição) de processado, e felicidade, como agentivo (condição) de feliz.
 - sobre a **inicialização** do estado descrito através de operações executadas pelos próprios objetos da classe relacionada a este estado. Exemplo: operar, como agentivo (inicialização) de operador, e nascer, como agentivo (inicialização) de nascido.
- **Télico.** Consiste na caracterização, através de operações, especificando funções ou propósitos dos objetos da classe relacionada ao estado descrito, enquanto estão neste estado. Exemplo: operar, como télico de operante.

3.1 Diferenças entre a Qualia Proposta e a Original

Alguns itens lexicais podem apresentar mais de uma Qualia, e elas podem ser de diferentes categorias gramaticais como, por exemplo, construção que é descrito [2] como o ato (SAB) de construir algo, ou uma edificação, ou seja, um objeto (SCC) construído. Uma descrição específica é possível porque há uma Qualia específica para cada caso. O quadro 2 mostra estas estruturas para o item lexical construção fazendo presente, mesmo que de forma indireta, a estrutura de eventos com subeventos processo (SAB) e estado (SCC) pretendida pela TLG.

<p><u>Construção</u> Qualia SAB Formal: <u>ato</u> Constitutivo: <u>início</u>, <u>duração</u>, <u>prazo</u>, <u>término</u>, <u>responsável</u>, ... Agentivo: <u>construtor</u> Télico: <u>construir</u> Qualia SCC Formal: <u>objeto</u> Constitutivo: <u>local</u>, <u>proprietário</u>, ... Agentivo-Criação: <u>construir</u></p>

Quadro 2. Qualia proposta para o item lexical construção como SAB e como SCC

O quadro 3 mostra a estrutura Qualia do item lexical livro, tratado como um SCC. No T-Lex, procura-se estender ao máximo a lista de valores dos papéis com o objetivo de alcançar a função do thesaurus, que é a de compor relacionamentos entre os itens lexicais. Com essa preocupação, no caso do papel Agentivo de livro, é alcançado outro objetivo descritivo: informar o fato de publicação, como objeto-físico, conter (ser suporte de) informação e, assim, atender o paradigma conceitual lexical, proposto na TLG, objeto-físico•informação.

<p><u>Livro</u> Qualia SCC Formal: <u>Publicação</u> Constitutivo: <u>capa</u>, <u>folha</u>, <u>caderno</u>, <u>título</u>, <u>texto</u>, <u>autor</u>, <u>figura</u>, <u>tabela</u>, <u>nota</u>, <u>índice</u>, <u>informação</u>,... Agentivo-Criação: <u>autor</u>, <u>escrever</u>, <u>editor</u>, <u>editar</u>, <u>gráfica</u>, <u>imprimir</u> (<u>título</u>, <u>texto</u>, <u>figura</u>, <u>tabela</u>, ...), ... Télico: <u>informar</u> (<u>leitor</u>)</p>

Quadro 3. Qualia proposta para o item lexical livro como SCC

Na quadro 4, podem ser observadas as estruturas Qualia do item lexical leitor, como SCC e como ADJ. Note-se que a descrição de livro é complementada pela de leitor, e que a associação de livro aos itens lexicais ler e leitura, por exemplo, é possível pela combinação das estruturas Qualia de livro e de leitor.

<p>leitor</p> <p>Qualia SCC</p> <p>Formal: <u>pessoa</u></p> <p>Agentivo-inicialização: <u>ler</u></p> <p>Télico: <u>ler</u></p> <p>Qualia ADJ</p> <p>Agentivo-condição: <u>leitura</u></p> <p>Agentivo-inicialização: <u>ler</u></p> <p>Télico: <u>ler</u></p>

Quadro 4. Qualia proposta para o item lexical leitor como SCC e como ADJ

4 IMPLEMENTAÇÃO DO T-LEX

Feita esta apresentação dos aspectos da aplicabilidade da estrutura Qualia, passamos a relatar como os mesmos foram implementados no presente trabalho.

Foi utilizado um corpus de teste com 7095 palavras, constituído por 34 resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Faculdade de Informática da PUCRS. Em média, os documentos possuem 208 palavras cada um e abordam assuntos da área de sistemas de informação.

O protótipo desenvolvido utiliza formulário html, para entrada e visualização dos dados, e programação CGI em ambiente Unix, com linguagem C, para o gerenciamento da base de dados.

Foram inseridas, até o momento, as descrições correspondentes: aos itens lexicais mais frequentes encontrados no corpus trabalhado; a todos os termos das consultas de teste da aplicação mencionada na seção 5; e, naturalmente, aos itens necessários para compor as descrições e que vão surgindo a cada inserção. As descrições dos itens lexicais, inseridas manualmente, baseiam-se nos verbetes contidos no dicionário Aurélio [2].

4.1 Linguagem de Descrição

No âmbito do presente trabalho foi desenvolvida uma linguagem de descrição para os valores dos papéis que compõem a Qualia de um item lexical. Esses papéis são preenchidos por descritores (outros itens lexicais) utilizando-se a seguinte notação:

Supertipo:Tipo

Supertipo é um item lexical utilizado para definir uma especificação genérica, de modo a precisar melhor um descritor, quando há ambigüidade por polissemia. Por exemplo, o supertipo ato em ato:construção identifica construção² como um ato e não como um efeito, entre outras coisas.

Operação(Argumento)

Argumento é um item lexical utilizado para identificar um SCC ou um SAB. Especifica um objeto direto ou indireto de um VRB, ou seja, de uma operação. Esta pode ser representada apenas pelo VRB ou por este seguido de seus argumentos. Por exemplo, em processar(linguagem), temos a operação processar e o argumento linguagem.

Conectores

São os operadores lógicos E e OU utilizados na listagem de itens lexicais de um papel. São representados por vírgula (,) e *pipe* (|), respectivamente.

Modificado.Modificador

O ponto é utilizado para viabilizar a representação de expressões, ligando um termo “modificado” a um ou mais termos “modificadores”. Por exemplo, podemos ter expressões com dois substantivos, como processamento de linguagem representada por processamento.linguagem, ou com um substantivo e um adjetivo, como processamento paralelo representada por processamento.paralelo, ou com um substantivo e um verbo, como computador processa representada por computador.processar(), ou com um verbo e um advérbio, como processar lentamente representada por processar().lento (o advérbio é, quando possível e necessário, modelado como um adjetivo). Observa-se que expressões com mais de dois termos podem ser utilizadas nas descrições. Por exemplo, processamento de linguagem natural seria representada por processamento.linguagem.natural.

² Construção significa “ato, efeito, modo ou arte de construir ...” [2].

–Hipônimo

É utilizado o hífen, como prefixo, para indicar um descritor como hipônimo (e não como hiperônimo) no papel formal de um item lexical. Quando isto ocorre, os papéis restantes da Qualia descrevem o hipônimo e não o item lexical principal. Desta forma viabilizam-se as descrições de expressões. Exemplo: –programação.concorrente (como valor do papel formal na descrição de programação). Neste caso, –programação.concorrente é descrita pelos papéis restantes desta Qualia, obrigando a existência de outra Qualia para descrever programação.

4.2 Codificação dos Itens Lexicais

Os itens lexicais no T-Lex são constituídos por bases e terminações, armazenadas separadamente, com códigos gerados automaticamente no momento da inclusão da base, em números decimais, e da terminação, em combinações de letras maiúsculas e/ou minúsculas.

Base

Corresponde aproximadamente ao conjunto formado por prefixo + radical, incluindo ou não a vogal temática. Por exemplo, a base auxili tem código 24.

Terminação

Corresponde aproximadamente ao conjunto formado por desinência + sufixo, incluindo ou não a vogal temática. Por exemplo, as terminações ar e ares formam uma classe com padrão ar e código E (neste caso, devido à ordem de inclusão, foi gerado um código em maiúscula), as terminações ar, ava, ei, amos, ... formam uma classe também com padrão ar, mas com código f, e as terminações ador, adores, adora e adoras formam uma classe com padrão ador e código d. O padrão é utilizado para visualização de uma base com uma classe de terminações.

Item lexical

Deve-se observar que, com esta codificação, construção e construções, por exemplo, são armazenados como um único item lexical. O mesmo ocorre com operador, operadores, operadora e operadoras, ou com o verbo falar e todas as suas flexões. Ou seja, os itens lexicais são armazenados em forma “canônica”, isto é, com gênero e número masculino singular, para substantivos e adjetivos, e com forma infinitiva, para os verbos. Utilizando os códigos das bases e das terminações, os itens lexicais auxiliar e auxiliares são armazenados pelo código 24E e são visualizados como auxiliar (base + padrão de terminação).

O uso destes códigos tem três objetivos:

- (i) economia de memória por armazenar itens lexicais (utilizando-se, em média, um número menor de caracteres);
- (ii) diminuição de ambigüidade na descrição. Por exemplo, o item lexical auxiliar utilizado como verbo numa descrição, é armazenado pelo código 24f (que representa tanto auxiliar como auxiliava, auxilie, auxiliamos, auxiliarei e todas as flexões do verbo auxiliar), e quando referenciado como substantivo ou adjetivo, é armazenado pelo código 24E (que representa auxiliar e auxiliares); e
- (iii) reconhecimento de ambigüidade na análise de termos. Usando o mesmo exemplo de (ii), quando se analisa o termo auxiliar, identifica-se a ambigüidade porque é possível “traduzí-lo” para dois códigos diferentes: como verbo (24f) e como substantivo ou adjetivo (24E).

5 DAS OPERAÇÕES GERATIVAS À EXPANSÃO DE CONSULTA

5.1 Operações Gerativas e Campo Lexical

O conceito de campo lexical, que será visto a seguir, é utilizado neste trabalho para, a partir de um dado termo, obter novos termos relacionados. Desse modo, tal conceito pode ser adotado, por exemplo, para expandir automaticamente a consulta originalmente feita em um sistema de RI.

O campo lexical é o conjunto dos lexemas que estão semanticamente relacionados paradigmática ou sintagmaticamente, dentro de um sistema lingüístico; é um subconjunto do léxico de uma língua [8][10].

Campo semântico é o conjunto de palavras constituído pela rede de relações tais como homonímia, sinonímia e polissemia [13].

O campo lexical (ou campo semântico) de um item lexical α pode ser obtido pelo conjunto dos itens lexicais gerados a partir das seguintes operações gerativas [3]: especialização, co-herança, associação, equivalência, decomposição e agregação.

especialização(a)

Obtém, através do papel formal onde α é um descritor, o conjunto de todos os itens lexicais descritos que são, assim, hipônimos de α . Exemplo:

especialização (dimensão) = { comprimento, altura, largura, ... }

co-herança(a)

Pesquisando a base de dados, obtém o conjunto de todos os itens lexicais com o mesmo hiperônimo de α no papel formal. Exemplo:

co-herança (comprimento) = { altura, largura, ... }

associação(a)

Pela leitura da descrição de α , obtém, através dos papéis agente e tético, o conjunto de todos os itens lexicais que são descritores nestes papéis e, assim, têm relação de implicatura/pressuposição [9] com α . Exemplo:

associação (processador) = { processamento, processar, ... }

equivalência(a)

Pesquisando a base de dados, obtém o conjunto de todos os itens lexicais β , tal que $\beta \in \text{co-herança}(\alpha)$, e $\text{associação}(\beta) \cup \text{associação}(\alpha)$ não seja vazio, sendo presumidos, assim, equivalentes a α . Exemplo:

equivalência (estudo) = { exame, análise, ... }

Observe-se que esta é uma operação configurável, sendo possível afetar o tamanho do conjunto resultante determinando o tamanho mínimo da interseção $\text{associação}(\beta) \cap \text{associação}(\alpha)$.

decomposição(a)

Pela leitura da descrição de α , obtém, através do papel constitutivo, o conjunto de todos os itens lexicais que são descritores neste papel e, assim, fazem parte de α . Exemplo:

decomposição (programa) = { estrutura de dados, função, procedimento }

agregação(a)

Pesquisando a base de dados, obtém, através do papel constitutivo onde α é um descritor, o conjunto de todos os itens lexicais descritos do qual α faz parte. Exemplo:

agregação (folha) = { árvore, livro, ... }

5.2 Termos Composicionais e Periféricos

Executando as operações descritas anteriormente sobre todos os termos de uma expressão em linguagem natural, é obtida uma série de conjuntos-resposta como resultado de cada operação sobre cada termo. As interseções desses conjuntos-resposta fornecem pistas sobre a semântica da expressão auxiliando, por exemplo, a resolução de ambigüidades. Com a aplicação de pesos maiores aos termos com maior frequência de ocorrência nos resultados das operações, decreta-se a maior importância semântica dos mesmos na interpretação composicional do significado da expressão.

Pode-se interpretar esses conjuntos-resposta como sendo compostos por termos composicionais e periféricos:

- Os composicionais, com maior peso, são aqueles termos encontrados em mais de um conjunto. Esses itens lexicais trazem indicações sobre o significado global da sentença analisada.
- Os outros, os periféricos, com menor peso, são ainda assim importantes pela probabilidade de ocorrerem no texto onde a sentença (mesmo com todos ou alguns de seus termos originais ausentes) refletiria o assunto tratado no todo ou em parte.

Com essas propriedades, as operações gerativas podem ser aplicadas, por exemplo, na expansão automática de consulta em sistemas de RI.

5.3 Expansão Automática de Consulta: Uma Aplicação Preliminar

Como exemplo de aplicação da geração do campo lexical, através de operações gerativas executadas sobre descrições de itens lexicais contidas no T-Lex, apresentamos resultados da avaliação comparativa entre mecanismos de busca de documentos em duas versões: com e sem expansão automática de consulta. Deve-se salientar que a implementação do thesaurus está em fase de protótipo e as operações gerativas estão sendo testadas ainda merecendo ajustes.

Nesta avaliação, foi utilizado um corpus de teste com resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Faculdade de Informática da PUCRS. Foram realizadas 24 consultas em cada uma das versões dos mecanismos de busca avaliados, e os resultados [5] de precisão e resposta foram calculados tomando-se os valores médios de precisão (nº de documentos relevantes recuperados / nº de documentos recuperados) e de resposta (nº de documentos relevantes recuperados / nº de documentos relevantes) para as 24 consultas, de acordo com os procedimentos adotados pela comunidade internacional de RI nas “Text Retrieval Conferences - TRECs” [17].

Na tabela 1 são apresentados os valores de precisão e de resposta obtidos. O valor de precisão considera o documento do topo da classificação nas duas versões de busca. O valor de resposta considera os quatro documentos recuperados como mais relevantes em cada caso. Nota-se que, nessas condições, a precisão da versão com expansão automática de consulta foi 8,6% superior, e a resposta 8,5% superior.

expansão automática de consulta	precisão (%)	resposta (%)
com	63	64
sem	58	59

Tabela 1. Valores de precisão e resposta obtidos na avaliação comparativa

6 CONSIDERAÇÕES FINAIS

O thesaurus idealizado e implementado permite uma série de relacionamentos lexicais, incluindo hiponímia, meronímia e implicatura/presuposição, de forma direta, através de descritores imediatos, ou indireta, através de descritores de descritores. Isso é possível pela execução das operações gerativas sobre as estruturas da base de dados.

A redução de ambigüidade lexical é obtida de duas maneiras. A primeira, através da codificação da base e da terminação que compõem os itens lexicais inseridos no thesaurus. A segunda, no tratamento de sentenças abordando, também, a capacidade composicional das palavras, é obtida pela análise da interseção dos campos lexicais de cada termo na aplicação das operações gerativas.

Embora outras aplicações sejam possíveis, o T-Lex foi idealizado como ferramenta auxiliar em sistemas de RI. Nesse sentido, é útil em tempo de indexação, como mecanismo de normalização semântica, e em tempo de busca, por exemplo, na expansão automática de consulta, como vimos.

REFERENCIAS BIBLIOGRÁFICAS

1. Cegalla, Domingos Paschoal. Minigramática da Língua Portuguesa. Cia. Editora Nacional, 1991. 496 p.
2. Ferreira, Aurélio B. de H. Dicionário Aurélio Eletrônico – Século XXI. Versão integral do Novo Dicionário da Língua Portuguesa – Século XXI, Rio de Janeiro: Nova Fronteira S.A. Lexikon Informática Ltda., versão 3.0, 1999.
3. Gonzalez, M. O Léxico Gerativo de Pustejovsky sob o Enfoque da Recuperação de Informações. Trabalho Individual I, Programa de Pós-Graduação em Ciência da Computação (PPGCC), Faculdade de Informática, PUCRS, maio 2000.
4. Gonzalez, M. Thesauri. Trabalho Individual II. Programa de Pós-Graduação em Ciência da Computação (PPGCC), Faculdade de Informática, PUCRS, maio 2001.
5. Joyce, T.; Needham, R.M. The Thesaurus Approach to Information Retrieval. American Documentation, 1958, V.9, pp.192-197. In: SPARCK JONES, K; WILLET, P. (editores). Readings in Information Retrieval. California: Morgan Kaufmann Publishers, Inc., 1997. pp. 15-20.
6. Van Der Laan, R. H.; Ferreira, G. I. S. Tesaurus e Terminologia. In: XIX Congr. Bras. de Bibliotec. e Documentação e III Congr. Latino-Americano de Bibliotec. e Documentação, Brasil, 2000. CD-ROM

7. Loukachevitch, N. V.; Salli, A. D.; Dobrov, B. V. Automatic Indexing Thesaurus Intended for Recognition of Lexical Cohesion in Texts. NLDB'99 – 4th Int. Conf. on Applications of Natural Language to Information Systems, 1999. OCG Schriftenreihe, Lecture Notes, v.129, pp.203-208.
8. Lyons, J. Semantics. Cambridge: Cambridge University Press, 1977. 2 v.
9. Pustejovsky, J. The Generative Lexicon. Cambridge: The MIT Press, 1995. 298 p.
10. Rehfeldt, Gládis Knak. Polissemia e Campo Semântico (estudo aplicativo aos verbos de movimento). Porto Alegre: EDURGS / FAPA / FAPCCA, 1980. 172 p.
11. Ruge, G. Combining Corpus Linguistics and Human Memory Models for Automatic Term Association. In: STRZALKOWSKI, Tomek. Natural Language Information Retrieval. Kluwer Academic Publishers, 1999. pp.75-98.
12. Rumbaugh, J.; Blaha, M.; Premerlani, W.; Eddy, F.; Lorenzen, W. Object-Oriented Modeling and Design. Englewood Cliffs: Prentice-Hall, 1991. 500 p.
13. Scapini, Isabel K. Associações Interlexicais: Contribuição para um Dicionário Remissivo. Dissertação de Mestrado. Porto Alegre: Instituto de Letras e Artes, PUCRS, 1997. 169 p.
14. Sparck-Jones, K. Synonymy and Semantic Classification. Edinburgh: Edinburgh University Press, 1986. 285 p.
15. Sparck-Jones, K.; Willet, P. (editores). Readings in Information Retrieval. California: Morgan Kaufmann Publishers, Inc., 1997.
16. Tálamo, M. De F. G. M.; Lara, M. L. G. De; Kobashi, N. Y.. Contribuição da Terminologia para Elaboração de Tesouros. Ciência da Informação, Brasília, 1992. V.21, n.3, pp.197-200.
17. Voorhees, E.M.; Harman, D. Overview of the Eighth Text Retrieval Conference (TREC-8). Capturado em dezembro de 2000. Disponível na Internet em <http://trec.nist.gov/pubs.html>.