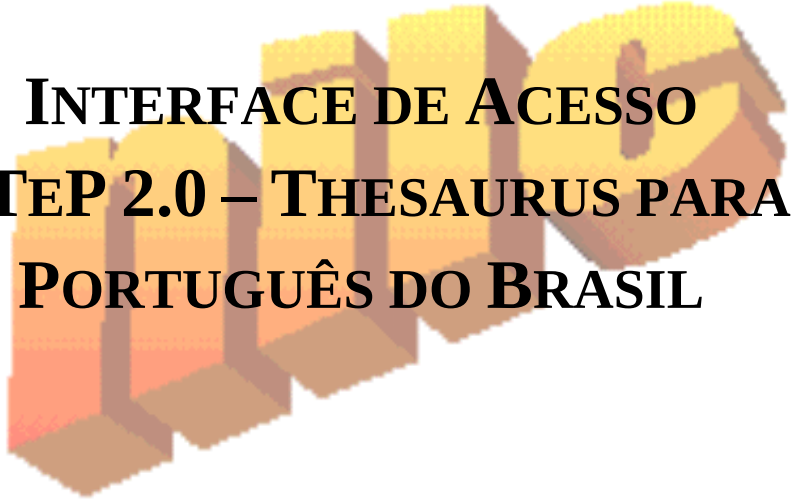
Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

INTERFACE DE ACESSO AO TEP 2.0 – THESAURUS PARA O PORTUGUÊS DO BRASIL



Erick Galani Maziero
Thiago Alexandre Salgueiro Pardo

NILC-TR-08-07
Junho, 2008

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

RESUMO

O Thesaurus para o Português do Brasil (TeP2.0) é uma base de dados que armazena conjuntos de palavras sinônimas e antônimas entre si. Tais relações se estabelecem dado o sentido, ou acepção, da palavra que se deseja consultar na base. Neste documento é apresentada uma interface web desenvolvida para facilitar a consulta ao TeP2.0. É abordada, também, a base de dados, com suas relações, armazenamento físico e processo de consulta.

Este trabalho contou com o apoio das agências de fomento à pesquisa FAPESP, CAPES e CNPq.

ÍNDICE

1. Introdução

Este documento apresenta o TeP 2.0, uma nova versão do TeP (Dias-da-Silva et al., 2000; Greghi, 2002; Greghi et al., 2002), um dicionário eletrônico de sinônimos e antônimos para o Português do Brasil. Para o referido thesaurus foi construída uma interface online para consulta à base de dados, que também é apresentada neste documento.

Um thesaurus é útil ao usuário que deseja, na escrita ou análise de um texto, ter opções de palavras sinônimas ou antônimas, por diversos motivos, dentre eles adequação comunicativa, precisão, correção ou aprendizagem. Além disso, é de vital importância para aplicativos de Processamento de Línguas Naturais, pois consiste em um primeiro passo para se lidar automaticamente com as palavras e seus sentidos.

Na próxima seção, apresenta-se a base de dados atual do TeP 2.0, com sua representação física. Na Seção 3, tratam-se das relações de sinônimo e antônimo do TeP. A interface de acesso online à base é apresentada na Seção 4, com o mecanismo de busca sendo delineado na Seção 5. Por fim, nas Seções 6 e 7, respectivamente, apresentam-se alguns dados da base e considerações finais.

2. A base de dados

2.1. Organização física

As palavras (ou verbetes) do TeP são agrupadas em conjuntos de sinônimos; estes conjuntos, comumente chamados de synsets (synonym set), são as unidades principais e, digamos, indivisíveis da base, a partir das quais se estabelecem todas as relações.

A atual base de dados do TeP 2.0 pode diferir da TeP original na forma em que os dados estão organizados fisicamente. Na Figura 1 é ilustrada a organização da base de dados.

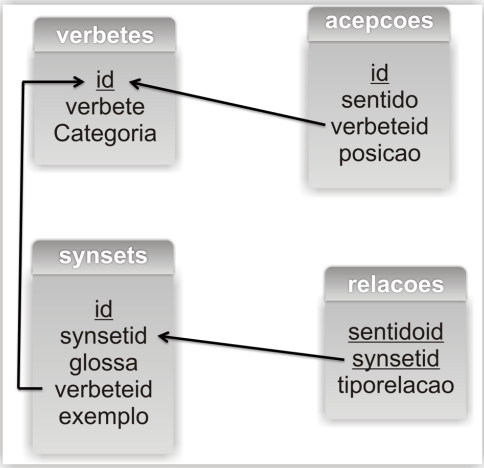


Figura 1: organização física da base de dados do TeP 2.0

Têm-se quatro tabelas: verbetes, acepções, synsets e relacoes. Na tabela **verbetes** encontra-se o verbete propriamente dito, com sua categoria gramatical. No TeP 2.0, os verbetes podem ser de basicamente 4 categorias gramaticais, a saber: verbo, substantivo, advérbio e adjetivo. Além de ter a opção de assumir a categoria "outra categoria" para verbetes que não se encaixem em alguma das 4 citadas anteriormente. A tabela **acepcoes** armazena o sentido do verbete, isto é, o significado da palavra. A tabela **synsets** indica quais os verbetes que pertencem a um mesmo conjunto de sinônimos. Os verbetes, indicados por *verbeteid*, que pertencem a um mesmo *synsetid*, formam um conjunto de sinônimos. Por fim, a tabela **relacoes** indica a relação que há entre dois conjuntos de verbetes. Essa relação pode ser sinônimo ou antônimo.

Essa organização reflete uma estrutura relacional dos dados, no sentido computacional do termo. Um exemplo de dados para o verbete "ascender" é exemplificado no próximo tópico.

2.2. Exemplo

O exemplo abaixo considera os registros físicos relacionados a uma busca pelo verbete "ascender":

verbetes = {'1553', 'ascender', '3'}
acepcoes = {'3240', 'chegar', '1553', '0'}
relacoes = {'3240', 'sinonimo', '86'}
synsets = {'247', '86', 'Glosa85', '1553', 'Affonso Moreira Penna, um dos mais expressivos juristas brasileiros, que há cem anos I]ascendia à Presidência da República.'}
synsets = {'248', '86', 'Glosa85', '1823', 'Mas a realidade I]avultou a seus olhos, e foi então que a alma tentou romper todos os elos e voar.'}
synsets = {'249', '86', 'Glosa85', '2377', 'Ano passado o volume apreendido não [chegou a 200 quilos, enquanto que só no primeiro semestre de 96 o total já é de 115 kg. '}
synsets = {'250', '86', 'Glosa85', '8224', 'O total dessas operações em 31/03/2001 I]montava R$ 136.881 mil.'}
synsets = {'251', '86', 'Glosa85', '10289', 'Dólar I]sobe a R$ 2,14 após leilão de títulos cambiais.'}
relacoes = {'3240', 'antonimo', '3120'}
synsets = {'13909', '3120', 'Glosa3119', '6652', 'x'}
acepcoes = {'3241', 'alçar-se', '1553', '1'}
relacoes = {'3241', 'sinonimo', '3071'}
synsets = {'14308', '3171', 'Glosa3170', '5141', 'x'}
synsets = {'14309', '3171', 'Glosa3170', '7331', 'x'}
synsets = {'14310', '3171', 'Glosa3170', '7500', 'x'}
synsets = {'14311', '3171', 'Glosa3170', '7607', 'x'}
synsets = {'14312', '3171', 'Glosa3170', '7627', 'x'}
synsets = {'14313', '3171', 'Glosa3170', '8132', 'x'}
synsets = {'14314', '3171', 'Glosa3170', '8172', 'x'}

relacoes = {'3241', 'antonimo', '625'}
synsets = {'1828', '625', 'Glosa624', '3597', 'x'}

No exemplo acima, pelas identações das linhas, podemos perceber que o verbete “ascender” possui duas acepções. Para as duas acepções, têm-se as duas relações: sinônimo e antônimo e, para cada relação, uma lista de verbetes que pertencem a um mesmo conjunto, que é sinônimo ou antônimo de cada acepção de “ascender”.

Na Seção 5 serão apresentados mais detalhes de como obter os conjuntos sinônimos a um dado verbete, o conjunto antônimo de cada acepção e exibição de frases-exemplo para o verbete pesquisado.

3. As relações na TeP 2.0

O Tep 2.0 conta atualmente com duas relações: a sinonímia entre palavras de um mesmo conjunto (synset) e a antonímia entre conjuntos de palavras.

3.1. Sinônimos

Sinônimo é uma relação entre duas ou mais palavras, em que os sentidos destas sejam idênticos ou semelhantes. Assim, *moto*, *motocicleta* e *motociclo* são palavras com o mesmo significado, usadas para designar o mesmo objeto, em geral. O advérbio *onde* pode ser substituído pelas expressões, *em que lugar* e *em qual lugar*.

3.2. Antônimos

Antônimo é o inverso do sinônimo; é a relação entre duas palavras, em que uma tem significado oposto, ou inverso, ao de outra. Desta forma, *subir* tem como antônimo a palavra *descer*.

No TeP 2.0, essa relação é estabelecida entre dois conjuntos de palavras sinônimas. Tal fato não impede o estabelecimento de antonímia entre duas palavras, bastando que estas estejam em dois conjuntos antônimos.

4. A interface web

4.1. A interface gráfica

Originalmente, na concepção do TeP, foi criada uma interface de acesso aos dados que se encontra em http://www.nilc.icmc.usp.br:8800/Diadorim/Thes.asp (endereço acessado em 03 de junho de 2008), que listava as possíveis acepções de dado verbete, e, para cada acepção, as palavras sinônimas e, se houvesse, os antônimos àquela acepção. A interface

original não é amigável e também não há a possibilidade de exibir possíveis exemplos dos verbetes procurados.

Na Interface do TeP2.0, adicionou-se uma interface mais amigável e intuitiva, com a opção de exibição de frase-exemplo ao verbete procurado. A Figura 2 mostra a tela inicial, com indicação de seus componentes e suas funções.

Figura 2: Elementos da interface gráfica disponível na web

Como exemplificação do uso da interface, realiza-se a busca pelo verbete *andar*. Para isto, deve-se digitar o verbete, em sua forma canônica (ou lematizada), na caixa de texto ao lado do botão “Buscar”. Visto que um verbete pode pertencer a mais de uma categoria gramatical, o usuário pode optar por especificar uma categoria aos conjuntos sinônimos retornados na busca. Caso não se queira especificar, pode-se deixar a opção “Todas”. Assim, não será realizada restrição sobre a categoria gramatical do verbete na busca. Na Figura 3, é ilustrado o procedimento de escolha da categoria.

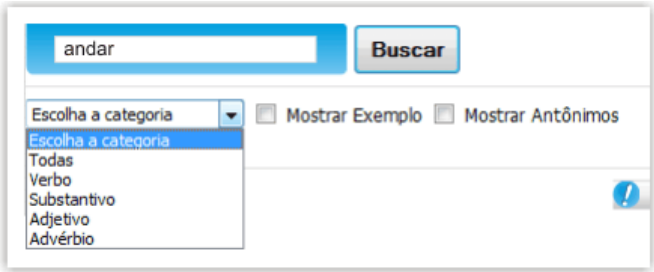


Figura 3: Escolhendo a categoria gramatical

Realizando a busca, são retornados conjuntos de sinônimos, como ilustrado na Figura 4, em que cada conjunto de palavras sinônimas é numerado seqüencialmente.

Figura 4: amostra de resultado de busca pelo verbete “andar”

Caso se deseje que na busca sejam exibidos os antônimos aos conjuntos listados, deve-se habilitar a opção “Mostrar Antônimos”, e realizar a busca novamente. Tem-se o resultado na Figura 5, em que, para cada conjunto retornado, exibi-se, quando há, um conjunto antônimo, rotulado por “Antônimos”.

Figura 5: amostra de resultado a busca anterior, com antônimos

Como mais uma opção de exibição de conteúdo na busca, pode-se exibir, quando houver, frase-exemplo ao verbete procurado. Basta habilitar a opção “Mostra Exemplo” e refazer a busca. Um exemplo encontra-se na Figura 6, em que o verbete, referente à busca, aparece sublinhado.

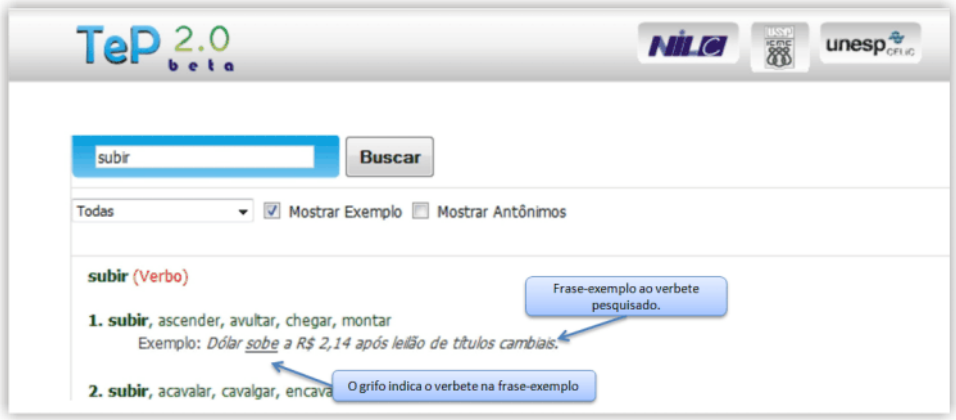


Figura 6: retorno da busca pelo verbete “subir” com opção de exibir exemplo

4.2. A base para download

Como mostrado na Figura 2, há a opção de fazer o download da base atual do TeP 2.0. Essa base consiste em um arquivo de texto simples, que contém todos os conjuntos de sinônimos do TeP 2.0 com as relações de antonímia entre os conjuntos.

O referido arquivo contém um conjunto de sinônimos por linha. Assim, as palavras que pertencem a uma mesma linha, pertencem a uma mesma acepção. No fim de cada linha há um número entre <> (quando a acepção tem algum conjunto de sinônimos antônimos à acepção), que indica o conjunto de sinônimos que é antônimo ao conjunto da presente linha. Assim tem-se o seguinte formato:

Numero_identificador_do_conjunto_de_sinonimos. [nome_da_categoria_gramatical_do_conjunto] {Conjunto de sinônimos separados por vírgula} <numero_identificador do conjunto de antônimos>

Abaixo há um exemplo:

3055. [Verbo] {desquietar, encabular, encafifar, grilar, impressionar, inquietar, intrigar, preocupar} <1307>

5. Sistema de busca na base de dados

Na segunda seção deste documento foi apresentado um exemplo de registros de dados referentes ao verbete *ascender*. Abaixo lista-se o processo de busca na base de dados atual do TeP 2.0 para o mesmo verbete.

Consulta 1:
Como a consulta é feita pelo verbete propriamente dito, procede-se inicialmente a uma busca na tabela *verbetes* pelo verbete digitado. Nesta consulta, obtém-se o identificador único do verbete, que neste caso é igual a ‘3’:

verbetes = {‘1553’, ‘ascender’, ‘3’}

Consulta 2:
Tendo o identificador do verbete, buscam-se todas as suas acepções. Neste caso, obtemos duas acepções:

acepcoes = {'3240', 'chegar', '1553', '0'}
acepcoes = {'3241', 'alçar-se', '1553', '1'}

Cada registro de acepção contém um campo de identificação único. Neste caso, tivemos o retorno de dois identificadores de acepção, o ‘3240’ e o ‘3241’.

Consulta 3:
Com posse dos identificadores da(s) acepção(ões), pesquisa-se na tabela *relacoes*, que indica, para cada acepção, as suas possíveis relações. Estas relações, como já dito, podem ser sinônimo e antônimo.
Para a acepção ‘3240’, têm-se as relações:

relacoes = {'3240', 'sinonimo', '86'}
relacoes = {'3240', 'antonimo', '3120'}

e para a acepção ‘3241’:

relacoes = {'3241', 'sinonimo', '3071'}
relacoes = {'3241', 'antonimo', '625'}

Desta forma, percebemos que, para as duas acepções do verbete *ascender*, temos conjuntos de sinônimos e antônimos. Pode acontecer na atual base de dados que um verbete tenha apenas sinônimos e não apresente conjuntos antônimos.
Cada registro da tabela *relacoes* contém um campo que indica o identificador do conjunto de sinônimos da relação, isto é, tem-se um conjunto de sinônimos que se relaciona com o verbete pela relação encontrada.
Para a relação de sinônimo da acepção identificada por ‘3241’, tem-se o identificador de conjunto de sinônimos ‘3071’. Para a relação de antônimos da acepção identificada por ‘3241’, tem-se o identificador de conjunto de sinônimos ‘625’.

Consulta 4:
Conhecendo os identificadores dos conjuntos acima citados, obtêm-se, pela tabela *synsets*, todos os identificadores dos verbetes que pertencem ao conjunto. Por exemplo, para o identificador ‘86’, da relação de sinônimo, da acepção ‘3240’ do verbete *ascender*, temos os seguintes registros:

synsets = {'247', '86', 'Glosa85', '1553', 'Affonso Moreira Penna, um dos mais expressivos juristas brasileiros, que há cem anos I]ascendia à Presidência da República.'}

synsets = {'248', '86', 'Glosa85', '1823', 'Mas a realidade I]avultou a seus olhos, e foi então que a alma tentou romper todos os elos e voar.'}

synsets = {'249', '86', 'Glosa85', '2377', 'Ano passado o volume apreendido não [chegou a 200 quilos, enquanto que só no primeiro semestre de 96 o total já é de 115 kg. '}

synsets = {'250', '86', 'Glosa85', '8224', 'O total dessas operações em 31/03/2001 I]montava R$ 136.881 mil.'}

synsets = {'251', '86', 'Glosa85', '10289', 'Dólar I]sobe a R$ 2,14 após leilão de títulos cambiais.'}

que contêm, no segundo valor de cada conjunto, o identificador do verbete. Este identificador é usado para uma outra busca na tabela verbetes, para obter o verbete propriamente dito. Vemos, também, que o último valor de cada conjunto, dos listados acima, contém uma frase exemplo ao verbete do conjunto de sinônimos.

6. Dados da base

O TeP 2.0 conta atualmente com 19.888 conjuntos de sinônimos, sendo que há 44.678 palavras no total, 2,5 palavras em média por conjunto de sinônimos e 18.163 relações de antonímia.

A base de dados TeP é feita manualmente por especialistas da área e encontra-se em constante ampliação e aprimoramento. É natural, portanto, haver erros e inconsistências nas informações armazenadas. Por exemplo, para o verbete *subir*, há a repetição da primeira acepção retornada pela interface, como mostrado na Figura 7.

Figura 7: exemplo de repetição de conjunto de sinônimos em busca

7. Considerações finais

O thesaurus para o Português do Brasil conta, como explicado e exemplificado neste documento, com as relações de sinônimo e antônimo. Trabalha-se para que a base de dados do TeP2.0 seja acrescentada de outras relações entre as palavras, tornando-se uma wordnet para o português do Brasil. Isto é possível, dado que a unidade básica de uma wordnet é a acepção ou conceito, que é como um nó na rede de relações (veja, por exemplo, Dias-da-Silva et al., 2006). A base do TeP2.0 armazena as palavras em acepções, bastando o acréscimo de relações, que vão além de sinônimos e antônimos.

O TeP 2.0 encontra-se disponível no link http://www.nilc.icmc.usp.br/tep2/.

Referências

Greghi, J.G.; Martins, R.T.; Nunes, M.G.V. (2002). Diadorim: a Lexical database for Brazilian Portuguese. In the *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1346-1350.

Greghi, J.G. (2002). *Projeto e desenvolvimento de uma base de dados lexicais do português*. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Dias-da-Silva, B.C.; Oliveira, M.F.; Moraes, H.R.; Hasegawa, R.; Amorim, D.; Paschoalino, C.; Nascimento, A.C.A. (2000). Construção de um Thesaurus Eletrônico para o Português do Brasil. In *Anais do V Encontro para o processamento computacional da Língua Portuguesa Escrita e Falada*. Atibaia, São Paulo, Brazil.

Dias-da-Silva, B.C.; Di Felippo, A.; Hasegawa, R. (2006). Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations. In the *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language*, pp. 120-130. Itatiaia, Rio de Janeiro, Brazil.