A Base de Dados Lexical e a Interface *Web* do TeP 2.0 – *Thesaurus* Eletrônico para o Português do Brasil

Erick G. Maziero, Thiago A. S. Pardo
NILC-ICMC – Universidade de São Paulo (USP)
P.O. Box 668 – 13560-970, São Carlos – SP, Brazil
+55 16 3373 9700
erickgm@grad.icmc.usp.br; taspardo@icmc.usp.br

Ariani Di Felippo, Bento C. Dias-da-Silva
NILC-FCL – Universidade Estadual Paulista
P.O. Box 174 – 14.800-901, Araraquara – SP, Brazil
+55 16 3301 6200
arianidf@uol.com.br; bento@fclar.unesp.br

ABSTRACT
In this paper, we describe the TeP 2.0 – Electronic *Thesaurus* for Brazilian Portuguese – which stores sets of synonym and antonym word forms. A *thesaurus,* as TeP 2.0, is an important resource for final users as much as for Natural Language Processing researches. Specifically, we present the lexical database and the *Web* interface of TeP 2.0.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language parsing and understanding*

General Terms
Languages and Theory

Keywords
Natural Language Processing; lexical resource; lexical database; *thesaurus*; Web nterface.

1. INTRODUÇÃO
Por um lado, um *thesaurus eletrônico* é útil ao usuário que deseja, na produção ou análise de textos, ter opções de escolha de sinônimos ou antônimos, por diversos motivos, dentre eles adequação comunicativa, precisão, correção ou aprendizagem. Por outro lado, é de vital importância para aplicativos de Processamento de Línguas Naturais (PLN), pois consiste em um primeiro passo para se lidar automaticamente com as palavras e seus significados, servindo de auxílio a diversas aplicações e tarefas, por exemplo, sumarização de textos, tradução automática, detecção de paráfrases, perguntas e respostas e recuperação e extração de informação, dentre várias outras.

O TeP 2.0, em especial, é a nova versão do TeP [1] [2], um dicionário eletrônico de sinônimos e antônimos para o português do Brasil (doravante, PB) disponibilizado por meio da base de dados lexical Diadorim[1] [3]. Do ponto de vista lingüístico, além da ampliação do número de entradas lexicais e dos conjuntos de sinônimos e antônimos em relação à versão inicial, o TeP 2.0 armazena determinadas informações provenientes da base da WordNet.Br (doravante, WN.Br) [4] [5], que está em pleno desenvolvimento. Do ponto de vista computacional, o TeP 2.0 está disponível para consulta e *download* via uma interface *Web*[2] amigável.

Para apresentar o TeP 2.0, este artigo foi organizado formalmente em 5 Seções. Na Seção 2, define-se o objeto "*thesaurus* eletrônico", posto que *thesaurus* é um termo empregado por diferentes especialistas para designar diferentes objetos. Na Seção 3, descreve-se a base lingüística do TeP 2.0. Na Seção 4, demonstra-se a interface *Web* de acesso aos dados. Por fim, na Seção 5, algumas considerações finais são feitas.

2. O OBJETO *THESAURUS* ELETRÔNICO
Buscando a precisão terminológica, define-se o objeto "*thesaurus* eletrônico", pois, como mencionado, o termo *thesaurus* é empregado por especialistas distintos para denominar objetos também distintos. Segundo [1] e [6], o TeP 2.0 pode ser definido como "um tipo específico de ferramenta de auxílio à expressão lingüística que pode ser integrado a processadores de textos". Assim definido, o TeP 2.0, quando acoplado a um processador de textos (por exemplo, o *Microsoft Word*), tem por finalidade oferecer ao usuário do PB a opção de sinônimos e antônimos que ele, por motivos diversos, como estilo, adequação comunicativa, precisão ou correção, queira substituir. Uma vez em formato eletrônico, o TeP 2.0 também pode ser empregado como recurso lingüístico-computacional em vários aplicações do PLN, como tradutores automáticos, sistemas de correção ortográfica e gramatical, sistemas de sumarização, entre outros. Nesse cenário, aliás, a construção de recursos lexicais que armazenam informações semântico-conceituais é premente, posto que tais informações são fundamentais para a compreensão e produção das línguas naturais. Para o PB, tais recursos ainda são escassos, apesar do reconhecido avanço dos estudos do PLN no Brasil.

3. A BASE DE DADOS DO TEP 2.0
Como mencionado, o TeP 2.0 pode ser visto como uma extensão do TeP, que fora desenvolvido segundo os pressupostos da

[1] http://www.nilc.icmc.usp.br/nilc/tools/intermed.htm
[2] http://www.nilc.icmc.usp.br/tep2/

WordNet de Princeton (doravante, WN.Pr) [7]. Conseqüentemente, o TeP 2.0 pauta-se especificamente: (i) na divisão das unidades lexicais nas categorias nome, verbo, adjetivo e advérbio; (ii) no construto *synset*; (iii) na definição de antonímia especificada pelos desenvolvedores da WN.Pr.

O ***synset*** (do inglês, *synonym set*) é o construto criado para designar a unidade básica de estruturação de uma rede *wordnet*, isto é, um conjunto de unidades lexicais sinônimas ou quase-sinônimas que permite ao falante inferir o conceito evocado pelas unidades. Em outras palavras, pode-se dizer que o *synset* é um conjunto de unidades lexicais de uma mesma categoria sintática que podem ser intercambiáveis em um determinado contexto, p.ex.: {bicycle, bike, wheel, cycle}. O *synset*, por definição, é construído de modo a codificar um único conceito lexicalizado por suas unidades constituintes.

O emprego do *synset* como construto representacional assume que o falante tem acesso aos conceitos expressos pelos itens lexicais de sua língua. A WN.Pr adota a noção de **sinonímia contextual** para a montagem de *synsets*. De acordo com essa noção de sinonímia, "duas unidades lexicais são sinônimas em um contexto C, se a substituição de uma pela outra em C não altera o valor de verdade de denotado por C" [7]. Assim, se o falante não conhece o significado de uma forma lexical, uma forma sinônima é suficiente para que ele identifique o conceito apropriado. Por exemplo, se o falante desconhece a forma x e essa forma é parte do *synset* y e o falante conhece as formas z e k desse *synset*, então, porque a forma desconhecida x é parte de y, o falante passa a ter acesso ao significado da forma x. Por exemplo, se o falante não conhece o significado da forma lexical *abafo*, ele pode acessar esse significado a partir do *synset* {abafo, agasalho}, se ele conhecer o significado da forma *agasalho*.

A outra relação armazenada do TeP é a **antonímia,** que engloba diferentes tipos de oposição semântica. São elas: **antonímia complementar**, que relaciona pares de itens lexicais contraditórios em que a afirmação do primeiro acarreta a negação do segundo e vice-versa, por exemplo: {vivo} e {morto}; **antonímia gradual**, que relaciona itens lexicais que denotam valores opostos em uma escala como, por exemplo, {pequeno} e {grande}; e **antonímia recíproca**, que relaciona pares de itens lexicais que se pressupõem mutuamente, sendo que a ocorrência do primeiro pressupõe a ocorrência do segundo como, por exemplo, {comprar} e {vender}.

Atualmente, o TeP 2.0 contém 19.888 conjuntos de sinônimos e 44.678 unidades lexicais, tendo a média de 2,5 unidades por conjunto de sinônimos. Quanto à antonímia, ressalta-se que há 4.276 relações entre os *synsets* da base do *thesaurus,* ou seja, aproximadamente 22% da base está relacionada por meio dessa relação. Além disso, para 253 unidades lexicais pertencentes à categoria dos verbos, o TeP 2.0 armazena uma frase-exemplo distinta para cada uma das unidades. A frase-exemplo fornece o contexto de uso mínimo do item lexical. O *thesaurus* armazena também uma glosa (ou seja, uma definição informal do conceito) para 6.648 *synsets,* todos eles constituídos por unidades da categoria dos verbos. Tanto as frases-exemplo quanto as glosas são informações provenientes da base da WN.Br [8], em desenvolvimento.

4. A INTERFACE *WEB*

O TeP 2.0, para consulta *Web*, é armazenado em um banco de dados relacional MySQL e as consultas são realizadas utilizando a linguagem PHP, com incorporação de outras linguagens de amplo uso na *Web* para confecção das páginas, como HTML, CSS e Javascript.

Além de acesso aos dados tradicionais, a interface *Web* do TeP 2.0 permite também a exibição de frases-exemplo para as formas léxicas procuradas e sugestões de outras formas caso a consulta não retorne resultado. A Figura 1 mostra a tela inicial da interface.

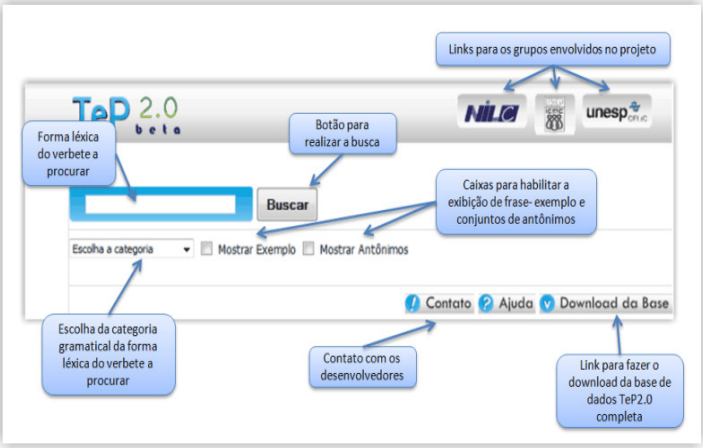

Figura 1. Tela inicial da interface *Web* do TeP 2.0

Como exemplificação do uso da interface, realiza-se a consulta da forma léxica do verbete *andar*. Para isto, deve-se digitar a forma canônica (ou lematizada) da palavra alvo da consulta na caixa de texto à esquerda do botão "Buscar". Dado que uma forma léxica pode pertencer a mais de uma categoria gramatical, o usuário pode optar por especificar uma categoria gramatical para os conjuntos sinônimos que deseja consultar. A opção "Todas" indica que não há especificação de categoria gramatical. Nesse caso, a interface exibirá todos os conjuntos de sinônimos ou *synsets* que contêm a forma léxica consultada e em todas as categorias registradas na base. A Figura 2 ilustra o procedimento de escolha da categoria gramatical.

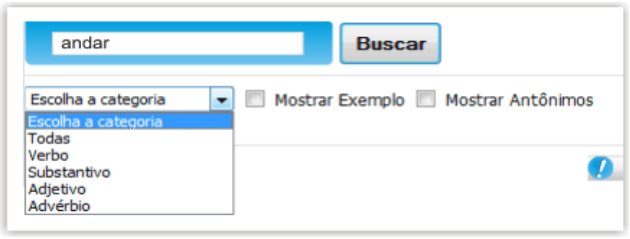

Figura 2. Seleção da categoria gramatical

A interface retorna para o usuário todos os conjuntos de sinônimos que contêm o verbo *andar*, como ilustra a Figura 3, em que cada conjunto é numerado seqüencialmente.

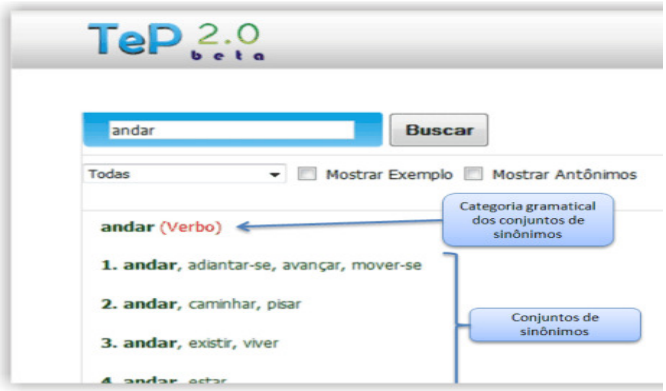

Figura 3. Os sinônimos do verbo *andar*

Habilitando-se a opção "Mostrar Antônimos" e fazendo-se uma nova consulta, a interface exibe, caso haja, os antônimos correspondentes aos conjuntos de sinônimos (Figura 4). Observa-se que o resultado da consulta é indicado pelo rótulo "Antônimos", exibido abaixo do conjunto de sinônimos ou *synset*.

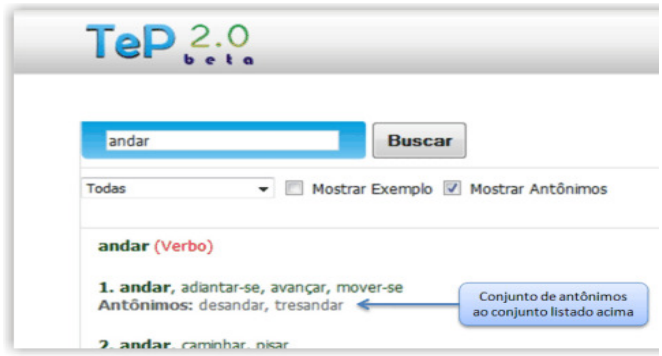

Figura 4. Os antônimos do *andar*

Como mais uma opção de consulta, pode-se exibir, quando houver, frases-exemplos para a forma consultada. Para essa opção, habilita-se o item "Mostra Exemplo" e realiza-se a consulta. A Figura 5 exemplifica essa funcionalidade da interface.

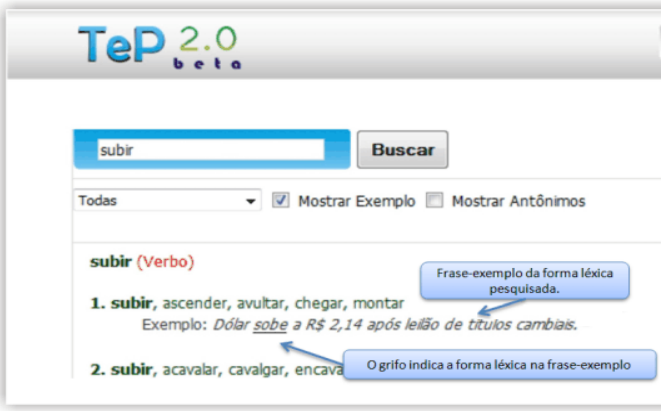

Figura 5. A frase-exemplo do verbo *subir*

5. CONSIDERAÇÕES FINAIS

Espera-se que o TeP 2.0, disponibilizado on-line para acesso pela comunidade, possa contribuir para o aprimoramento de aplicações de PLN existentes e para o desenvolvimento de aplicações novas. O conjunto completo de dados do TeP 2.0 pode ser salvo pelo usuário e incorporado nas mais diversas aplicações. No futuro, pretende-se estender a base de dados e a interface *Web*, possibilitando que estas suportem outros dados (p.ex.: relações de hiponímia) provenientes da WN.Br, além das glosas e frases-exemplos, que já foram inseridas no TeP 2.0.

6. REFERÊNCIAS

[1] Dias-da-Silva, B.C.; Oliveira, M.F.; Moraes, H.R.; Hasegawa, R.; Amorim, D.; Paschoalino, C. and 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.

[2] Oliveira, M.F. and Dias-da-Silva, B.C. 2006. A construção do thesaurus eletrônico para o português do Brasil (TeP) – pressupostos teórico-metodológicos. In Dias-da-Silva, B.C. e Longo, B.N.O. (Orgs), A construção de dicionários e bases de conhecimento lexical. Série Trilhas Lingüísticas, N. 9, pp. 187-208. Araraquara: Laboratório Editorial FCL/UNESP; São Paulo: Cultura Acadêmica Editora.

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

[4] Di Felippo, A. and Dias-da-Silva, B.C. 2007. Towards an automatic strategy for acquiring the WordNet.Br hierarchical relations. In Proceedings of the 5th Workshop in Information and Human Language Technology. Rio de Janeiro.

[5] Dias-da-Silva, B.C.; Di Felippo, A. and Nunes, M.G.V. 2008. The automatic mapping of Princeton WordNet lexical-conceptual relations onto the Brazilian Portuguese WordNet database. In Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco.

[6] Oliveira, M.F. 2002. Pressupostos teórico-metodológicos para a elaboração da base lexical de um thesaurus eletrônico. Dissertação de Mestrado. Araraquara: FCL – UNESP.

[7] Fellbaum, C. 1998. *WordNet: an electronic lexical database*. Ca., MA: MIT Press.

[8] Dias-da-Silva, B.C.; Di Felippo, A. and Hasegawa, R. 2006. Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations. In Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language (Lecture Notes in Computer Science 3960), pp. 120-130. Itatiaia, Rio de Janeiro, Brazil.