

Portal Min@s: Uma Ferramenta de Apoio ao Processamento de Córpus de Propósito Geral

Arnaldo Candido Junior^{1,3}, Thiago Lima Vieira², Marcel Serikawa², Matheus Antônio Ribeiro Silva², Régis Zangirolami², Sandra Maria Aluísio³

1. Secretaria de Educação Profissional e Graduação Tecnológica
Universidade Tecnológica Federal do Paraná, Medianeira, PR

2. Departamento de Computação
Universidade Federal de São Carlos, São Carlos, SP

3. Núcleo Interinstitucional de Linguística Computacional
Instituto de Ciências Matemáticas, Universidade de São Paulo, São Carlos, SP

{arnaldoc, sandra} at icmc.usp.br, {lima.vieira.thiago, marcel.serikawa, regismz} at gmail.com, mateusmoro at hotmail.com

***Abstract.** This paper presents Portal Min@s, a general web-based corpus processing tool. Many corpus processing tools available focus on specific tasks, such as lexicography or translation. Portal, on the other hand, took the challenge of being a general purpose corpus processing tool which deals with different types of corpus, languages and linguistic annotations. We present the features provided by this tool and compare it with two other alternatives.*

***Resumo.** Este artigo apresenta a ferramenta Portal Min@s, criada para apoiar a tarefa de processamento de córpis. Enquanto muitas ferramentas disponíveis focam em pesquisas específicas como lexicografia ou tradução, o Portal fornecendo recursos para tarefas mais gerais, processando córpis com diferentes propósitos, anotação e estruturação. Os recursos disponibilizados são detalhados e comparados com duas ferramentas similares.*

1. Introdução

Acompanhado o crescimento da linguística de córpis¹, uma grande quantidade de córpis foram compilados e disponibilizados para pesquisa linguística e para a criação de ferramentas de Processamento de Línguas Naturais (PLN). A maioria demanda ferramentas de processamento robustas devido ao seu tamanho. Em resposta a essa demanda, o número de ferramentas para processamento de córpis para apoiar os projetistas de córpis também tem aumentado. Muitas dessas ferramentas focam em córpis específicos (por exemplo, anotados ou não anotados) ou em tarefas específicas da linguística de córpis (por exemplo, tradução ou lexicografia). Nesse contexto, foi proposta a criação da ferramenta Web Portal Min@s² para uso em tarefas da linguística

¹ Neste trabalho optou-se pela grafia aportuguesada da palavra “córpus/corpora”.

² <http://fw.nilc.icmc.usp.br:12480/portal/>

de c3rpus. A ferramenta 3 livre e disponibilizada publicamente para todas as institui33es interessadas em seu uso.

Este trabalho 3 organizado como segue. A Se333o 2 apresenta uma vis33o geral do dos recursos e funcionalidades do Portal Min@s. A Se333o 3 compara o Portal com trabalhos relacionados. Por fim, a Se333o 4 apresenta as conclus33es do artigo.

2. Detalhamento do Portal Min@s

2.1. Projeto e Implementa333o

A primeira quest33o de projeto analisada foram as tecnologias a serem utilizadas no desenvolvimento do Portal Min@s. O Portal foi projetado em linguagem Java³ para ambiente Web, demandando o servidor Apache Tomcat⁴. Ao importar um c3rpus, os tokens de cada texto s3o armazenados em um banco de dados PostgreSQL⁵. As tarefas de tokeniza333o, segmenta333o sentencial e anota333o morfossint33tica s3o realizadas com o apoio da biblioteca OpenNLP [Baldrige, 2005]. A gera333o de n-gramas 3 baseada na ferramenta Ngram Statistics Package (NSP)⁶, enquanto que a extra333o de palavras-chave 3 feita pelo m3todo LDA (Latent Dirichet Allocation) [Blei, 2012]. O alinhamento autom3tico de lexemas para c3rpus paralelos 3 baseado na biblioteca Giza++ [Och, 2003], complementado com o TCAIign [Caseli, 2004]. A tarefa de lematiza333o 3 feita com base na sa3da do etiquetador morfossint33tico aliada aos dicion33rios DELA (Dictionnaire Electronique du LADL – Dicion33rio Eletr33nico do LADL) do Unitex [Paumier, 2006]. Por fim, a ferramenta GNU Aspell⁷ 3 aplicada para corre333o ortogr33fica autom3tica.

O Portal Min@s 3 utilizado para armazenar diversos c3rpus. Considerando sua natureza de acessos paralelos e a exist33ncia de c3rpus com milh3es de palavras (n3o h3a um limite m3ximo para o tamanho do c3rpus), efici33ncia e desempenho s3o fatores cr3ticos. Para lidar com essa quest33o, duas decis33es de projeto foram tomadas: (i) o uso de uma fila de tarefas de pr3-processamento e importa333o de c3rpus e (ii) o uso de bancos de dados para agilizar a recupera333o de dados, seguindo as recomenda333es de Davies [2005, 2009]. A fila reduz problemas de lentid3o de acesso durante a importa333o de grandes c3rpus. O tempo de importa333o pode variar de acordo com o tamanho do c3rpus e com os pr3-processadores escolhidos, demandando poucas horas no caso mais comum. Observa-se que usu3rios podem acessar c3rpus que est3o sendo importados.

Para otimizar as buscas, o Portal Min@s faz uso da estrutura de indexa333o madura e eficiente oferecida pelos Sistemas Gerenciadores de Banco de Dados. A estrutura atual do banco conta com 40 tabelas. A tabela mais importante se chama *word* e 3 utilizada para armazenar lexemas e outros *tokens*. Buscas por uma 3nica palavra s3o simples e diretas. Buscas por sequ3ncias de palavras s3o recuperadas por meio de *joins* da tabela *words* consigo mesma, conforme exemplificado no Quadro 1 para a busca por “vinho tinto”.

3 http://www.java.com/pt_BR/

4 <http://tomcat.apache.org/>

5 <http://www.postgresql.org/>

6 <https://metacpan.org/release/Text-NSP>

7 <http://aspell.net>

Quadro 1. A busca por palavras compostas no Portal Min@s

```
select * from words as w1, words as w2
where w1.word = "vinho"
and w2.word = "tinto"
and w2.postion = w1.position + 1
```

2.2. Principais Funcionalidades

Após a importação de um córpus, diversas funcionalidades são disponibilizadas ao usuário. O principal módulo é o concordanciador, sendo disponibilizado em duas versões: monolíngue e multilíngue. De acordo com o grau de anotação do córpus, diversas buscas podem ser realizadas como fragmentos de lexemas, informações morfo sintáticas como “pronomes seguidos de flexões do verbo ser” e informações no cabeçalho dos textos (como autor e editora). Buscas por palavras normalizadas, por exemplo, grafia atualizada para textos históricos, também são disponibilizadas.

Além do concordanciador, diversos outros módulos são disponibilizados. O módulo para alinhamento automático e semiautomático permite gerenciar córpuses multilíngues. Os módulos de estatística e frequências oferecem listagens de tokens/types, n-gramas (bigramas ou trigramas) mais frequentes e colocações. O módulo de palavras-chave oferece a extração de candidatas a palavras-chave através do método LDA. O módulo de edição de anotações foi inspirado na ferramenta Brat⁸ e permite anotar n-gramas e estabelecer relações binárias entre eles. Por fim, o módulo de córpus multimodais é um recurso simples para arquivamento de informações não textuais como transcrições fonéticas e imagens (os arquivos são apenas armazenados, sem manipulação direta dentro do Portal).

Usuários do Portal Min@s contam com uma série de recursos gerenciais extras para importar e controlar o acesso aos córpuses armazenados na ferramenta. O módulo de importação de textos é responsável por aplicar uma série de pré-processadores, tais como tokenização, lematização, alinhamento, dentre outros. Os módulos de gerenciamento permitem administrar córpuses, subcórpus e textos. Por meio deles, córpuses podem ser marcados como públicos ou privados e suas políticas de acesso como *copyright* e termos de uso são definidas. O submódulo para gerenciamento de etiquetas permite administrar diferentes categorias, incluindo etiquetas do córpus, dos textos (por exemplo, autor), de n-gramas (por exemplo, funções sintáticas), de seções do texto (notas de rodapé) e de formatação. Por fim, o módulo de gerenciamento de usuários permite o cadastro de usuários para acessar o Portal. Cinco perfis de usuários, com diferentes níveis de acesso são fornecidos: administrador, coordenador, colaborador, usuário regular e visitante.

3. Comparativo com Ferramentas Relacionadas

O comparativo desta seção segue a ISO 9126 [ISO, 1994], com foco no item funcionalidade, sendo baseado em duas populares ferramentas livres: Unitex⁹ [Paumier, 2006] e Philologic¹⁰ [University of Chicago, 2007]. Outros comparativos mais

⁸ <http://brat.nlplab.org/>

⁹ <http://www-igm.univ-mlv.fr/~unitex/>

¹⁰ <http://philologic.uchicago.edu/>

detalhados são realizados por Schulze et al. [1994], Santos & Ranchhod [2002] e Rayson [2002].

Unitex é um sistema de processamento de córpus baseado na teoria dos autômatos. Por se tratar de uma ferramenta em Java, é altamente portátil. Os recursos oferecidos pela ferramenta são agrupados em quatro funcionalidades principais: (a) autômatos, usados para criação de dicionários, buscas e transformações nos textos; (b) dicionários de apoio, utilizados, entre outras tarefas, para flexionar palavras automaticamente (alguns dos quais utilizados no Portal Min@s); (c) listagem de frequências; e (d) um concordanciador baseado em dicionários e autômatos. O Unitex oferece buscas baseadas em lemas e classes gramaticais, porém sem a eliminação de ambiguidade. Outra limitação é que apenas um texto ou córpus pode ser aberto de cada vez.

Philologic é um conjunto de ferramentas para processar córpus. Como o Portal Min@s, também é uma ferramenta Web capaz de atender a diversos usuários simultaneamente. As funcionalidades oferecidas pelo Philologic podem ser agrupadas em três grandes grupos: concordâncias, frequências e colocações e gerenciamento de subcórpus. Adicionalmente, a ferramenta oferece recursos para córpus multimodais de forma similar ao Portal. Textos devem seguir o padrão TEI Lite (Text Encoding Initiative Lite), mas podem ser personalizados até um certo limite. Um recurso semelhante à normalização ortográfica do Portal é utilizado para córpus históricos ou com erros de grafia através da ferramenta AGREP¹¹. De forma similar ao Portal, permite que as concordâncias sejam refinadas por parâmetros bibliográficos, fornecidos pelo cabeçalho TEI em cada texto. Assim como o Portal, é de difícil instalação por demandar um servidor Web e possuir diversas dependências.

4. Conclusões

Este trabalho apresentou a ferramenta Portal Min@s, em fase final de desenvolvimento e com diversos recursos para apoiar diferentes perfis de pesquisas em linguística de córpus em diferentes tipos de córpus. Atualmente, o Português, o Espanhol e o Inglês são processados, estando a introdução de recursos para outras línguas em andamento. O Portal está sendo testado sobre 10 córpus de tipos variados, incluindo jornalísticos, históricos, literários paralelos e de transcrições fonéticas, alguns dos quais em fase de importação. A ferramenta e seu código fonte estão disponíveis em <<https://bitbucket.org/portalminas/portal-minas/>>. Mais detalhes podem ser encontrados em Candido Junior et al. [2015]. Trabalhos futuros incluem um comparativo mais detalhado do Portal Min@s com ferramentas existentes para processamento de córpus, incluindo tempos para importação e acesso a córpus nas diversas ferramentas.

Agradecimentos

Agradecemos a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo financiamento do projeto. Agradecemos aos pesquisadores idealizadores e parceiros da Faculdade de Letras da Universidade Federal de Minas Gerais, cujo empenho possibilitou o desenvolvimento do Portal.

¹¹ <http://www.tgries.de/agrep/>

Referencias

- Baldrige, J. The Opennlp Project. 2005. Disponível em: <<http://opennlp.apache.org/index.html>>. Acesso em: Jun. 2014.
- Blei, D. M. Probabilistic Topic Models: Surveying a suite of algorithms that offer a solution to managing large documents archives. *Communications of the ACM*, s.l., n. 55, p. 77-84, 2012.
- Candido Junior, A.; Vieira, T. L.; Serikawa, M.; Silva, M. A. R.; Zangirolami, R.; Aluísio, S. M. *Portal Min@s: Uma Ferramenta Geral de Apoio ao Processamento de Córpus*. Série de Relatórios do NILC. NILC-TR-15-03, Agosto 2015, 11p.
- Caseli, H.d.M., Silva, A.M.d.P., Nunes, M.d.G.V.: Evaluation of methods for sen-tence and lexical alignment of brazilian portuguese and english parallel texts. In: *Advances in Artificial Intelligence (SBIA 2004)*, Lecture Notes in Computer Science, vol. 3171, pp. 184–193, 2004.
- Davies, M. "The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation". *International Journal of Corpus Linguistics* 10: 301-28, 2005.
- Davies, M. "Relational databases as a robust architecture for the analysis of word frequency". In *What's in a Wordlist?: In Investigating Word Frequency and Keyword Extraction*, ed. Dawn Archer. London: Ashgate. 53-68, 2009.
- ISO/IEC. 1994. ISO 9126: The Standard of Reference. 1994. Disponível em <<http://www.cse.dcu.ie/essiscope/sm2/9126ref.html>> (acessado Abril, 2015).
- Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. vol. 29, pp. 19–51. Association for Computational Linguistics, 2003.
- Paumier, S. Unitex 1.2: User Manual. 2006. Disponível em <<http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf>> (acessado Abril, 2015).
- PYYSAALO, Sampo et al. Brat Rapid Annotation Tool. 2012. Disponível <<http://brat.nlplab.org>> (acessado Maio, 2015).
- Rayson, P. E. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PHD Thesis. Lancaster University, 2002.
- Santos, D., and E. Ranchhod. 2002. Ambientes de processamento de corpora em português: comparação entre dois sistemas. In: *PROPOR'99*: Evora, 2002.
- Schulze, B. M. et al. Comparative State-of-the-Art Survey and Assessment Study of General Interest Corpus-oriented Tools. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1994.
- University of Chicago. PhiloLogic User Manual. 2007. Disponível em <<http://philologic.uchicago.edu/manual.php>> (acessado Abril, 2015).