Introdução ao PANTERA

Diana Santos, POR4103

Versão 1, 24 de novembro de 2014

1 O corpo PANTERA

O PANTERA é um corpo paralelo desenvolvido na Linguateca com o apoio da Universidade de Oslo, acessível de `http://www.linguateca.pt/PANTERA/`. Encontra-se em desenvolvimento e pretende incluir excertos de todas as obras traduzidas entre as línguas portuguesa e norueguesa.

Por essa razão, inclui também uma base de dados com essa informação, em constante atualização, e cuja interface, neste momento, é feita apenas através da página `http://www.linguateca.pt/PANTERA/ListaPANTERA.html`.

Para evitar a necessidade de pedir autorização aos diversos autores, tradutores e editores envolvidos, o que tornaria o projeto inviável em termos de trabalho e de possibilidade de disponibilização posterior, usamos apenas um excerto das obras, de tamanho entre dez a quinze páginas – exceto nos casos em que não haja direitos de autor envolvidos ou já tenhamos autorização para tal.

Embora a maioria das obras traduzidas entre as duas línguas sejam do foro literário, quaisquer textos de que tenhamos conhecimento serão incluídos, visto que um dos objetivos do PANTERA é propiciar os estudos contrastivos e de tradução.

2 Passos envolvidos na criação do PANTERA

Depois de identificar a existência de uma (ou mais) traduções de uma dada obra, é preciso ter acesso físico ao texto e sua tradução, quer através de formato eletrónico já existente (como é o caso dos textos antigos, ou de prévia compilação em outros corpos), quer através do acesso ao livro publicado.

Nesse caso, é preciso recorrer ao ROC – reconhecimento ótico de carateres – e efetuar a sua posterior revisão humana.

Depois disso os textos são alinhados automaticamente com as ferramentas associadas ao ambiente DISPARA [2], e o alinhamento é revisto numa fase posterior.

Em seguida os textos do lado português são analisados como todos os outros do projeto AC/DC [4], usando o PALAVRAS [1] e o resto da anotação semântica já existente [5]. O processo é o mesmo que para o CorTrad [6].

Infelizmente, ainda não tivemos tempo para usar um analisador sintático do norueguês, por isso o lado norueguês ainda não está analisado.

3 Uso do PANTERA

A primeira coisa importante é indicar que se pode procurar por qualquer campo, mas por omissão o campo é o das palavras/formas (chamadas `word`) que se encontram no texto.

Por isso, é possível procurar

```
comeu
```

que dá todos os casos da palavra *comeu* no PANTERA, ou

```
[lema="comer"]
```

que apresenta todos os casos do verbo comer (com lema *comer*) (mas veja-se abaixo para os casos de enclíticos ou mesoclíticos).

Quando se quiser procurar no par original com procuras simultâneas nas duas línguas, ou melhor com procuras que também selecionam com base na "outra" língua, usa-se as duas caixas de procura simultaneamente. Por exemplo

```
[lema="casar"]  [word="gift.*"]
```

procura os casos em que em português existe uma forma do verbo *casar*, e em norueguês uma palavra iniciada pelas letras `gift`, enquanto que a seguinte procura

```
[word="gift.*"] [lema="casar"]
```

encontra os casos em que em norueguês existe uma palavra iniciada por `gift`, e em português uma forma do verbo *casar*

Visto que todos os textos estão emparelhados, neste caso as duas procuras darão resultados idênticos (apenas com os lados das línguas trocados), mas é preciso sublinhar que, se estivéssemos interessados na direção da tradução, ou seja, nos casos em que *casar* é traduzido por *gift.** ou vice-versa as procuras, e os resultados, seriam distintos.

```
[lema="casar" & oritrad="ori"]  [word="gift.*"]
```

`oritrad` é uma informação que indica se o texto em questão é original ou traduzido. Nesta procura estou a restringir a procura a textos em português, traduzidos para o norueguês.

Outra coisa que se pode procurar no PANTERA é a distribuição. Por exemplo, em vez de procurar as próprias concordâncias, poderíamos estar interessados apenas em quantas vezes apareciam em texto traduzirdo ou original. Ness caso deveríamos escolher a opção Distribuição `Original/Traduzido`.

Mas a distribuição pode ser feita por todas as formas em que uma dada procura pode variar. Podemos pedir distribuição por `Fontes`, ou seja, pelas obras em que a nossa procura encontrou algo; ou por variante da língua, `Variante do português` ou `Variante do norueguês`.

Além disso, podemos estar interessados na distribuição das próprias formas (no caso de termos pedido algo que tenha várias formas), ou – apenas para o caso do português, por enquanto – distribuição dos lemas (`Lemas` e da categoria gramatical `PoS`. Eventualmente poder-se-á pedir a distribuição de mais informação que for útil para a análise das traduções, que ainda não se encontra disponível através da interface.

Outra possibilidade é empregar expressões de procura que se refiram a mais de uma palavra, por exemplo sequências de palavras

```
"til" "og" "med"
```

não necessariamente contíguas, como no seguinte caso em que estamos à procura da locução *ir* (ou *vir*) *embora*:

```
[lema="ir|vir"] []* "embora"
```

Finalmente, também é possível fazer perguntas negativas referentes ao texto alinhado:

```
gammel !velho
```

que significa casos em que haja a palavra *gammel* no lado esquerdo, mas não haja a palavra *velho* no direito. Reparem que isto é diferente de pedir

```
gammel [word!="velho"]
```

visto que esta procura é satisfeita por casos em que haja *gammel* do lado esquerdo e uma palavra qualquer que seja diferente de *velho* no lado direito, o que será sempre o caso.

Para explicações mais cabais do tipo de sintaxe usada no PANTERA, veja-se, além disso, o texto [3].

Referências

[1] Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento, Aarhus University, Aarhus University Press, November de 2000.

[2] Diana Santos. DISPARA, a system for distributing parallel corpora on the Web. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, Faro, Portugal, 23-26 de Junho de 2002. p. 209–218.

[3] Diana Santos. A sintaxe do AC/DC: apresentação do CWB e das opções tomadas. 2012. `http://www.linguateca.pt/Diana/download/instrACDC.pdf`.

[4] Diana Santos e Eckhard Bick. Providing Internet access to Portuguese corpora: the AC/DC project. Em Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis e Gregory Stainhauer, editores, *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, 31 de Maio - 2 de Junho de 2000. p. 205–210.

[5] Diana Santos e Cristina Mota. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. Em Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner e Daniel Tapias, editores, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 17-23 de Maio de 2010. p. 1437–1444.

[6] Stella O. E. Tagnin, Elisa Duarte Teixeira e Diana Santos. CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica*, 4:314–323, 2009.