Documentação de corpus - Corpus Histórico do Português Tycho Brahe

Laila Mota
Universidade Federal da Bahia – Instituto de Computação
laila.pereira@ufba.br

Resumo

Este artigo apresenta a documentação de toda a manipulação realizada no Corpus Histórico do português Tycho Brahe, uma coleção de textos escritos em português que abrange um período crucial da história da língua, utilizado treinamento de modelo para análise de evolução semântica. O corpus é composto por documentos que vão do século XIV ao século XX, período em que ocorreram importantes transformações linguísticas na língua portuguesa. A documentação abrange informações sobre a origem e a autenticidade dos textos, bem como a sua transcrição, anotação e organização. Além disso, são discutidas as questões metodológicas envolvidas na construção do corpus, incluindo a seleção de textos representativos, a padronização da ortografia e os procedimentos realizados para padronização dos textos. A documentação do corpus histórico do português Tycho Brahe oferece uma base sólida para estudos diacrônicos da língua portuguesa, possibilitando análises detalhadas das mudanças linguísticas ocorridas nesse período e contribuindo para o avanço do conhecimento sobre a evolução da língua.

Palavras chave

análise diacrônica, documentação, corpora, processamento de linguagem natural

1 Introdução

2 O Corpus

O Corpus Histórico do Português Tycho Brahe (Galves, 2018) é uma valiosa coleção de textos escritos em português escritos por autores nascidos entre 1380 e 1978. Esse período marca uma época de grande relevância para a evolução do português, marcada por transformações linguísticas significativas. A constituição desse corpus é de extrema importância para o estudo diacrônico da língua portuguesa, permitindo uma análise aprofundada das mudanças linguísticas ocorridas nesse período.

Trata-se de um corpus eletrônico anotado, composto em sua maioria de textos em português europeu e conta com textos em português brasileiro. Atualmente são 88 textos (3.544.628 palavras) que estão disponíveis para pesquisa livre, com um sistema de anotação linguística em duas etapas:

1. anotação morfológica (aplicada em 58 textos, num total de 2.280.819 palavras);
2. anotação sintática (aplicada em 27 textos, num total de 1.234.323 palavras).

A importância desse corpus reside no fato de oferecer subsídios para estudos da língua portuguesa no que diz respeito à uma visão detalhada das mudanças que ocorreram no vocabulário, na gramática, na sintaxe e nas expressões idiomáticas ao longo dos séculos. Através da análise comparativa desses documentos, é possível identificar e estudar as transformações linguísticas que moldaram o português ao longo do tempo.

A construção do corpus é resultado de pesquisas iniciadas em 1998 e contou com a colaboração de pesquisadores de diversas instituições como o Instituto de Estudos da Linguagem da Universidade Estadual de Campinas (IEL-UNICAMP), a Universidade de Lisboa e o Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP).

O acesso ao corpus encontra-se disponível para pesquisadores, gratuitamente, para fins acadêmicos e pedagógicos, através do site do corpus e pode ser feito via arquivos `txt` ou `xml` que podem ser baixados diretamente do site.

Os textos foram editados seguindo a sua classificação de acordo com sua natureza:

- Texto-Fonte com grafia preservada: nestes casos, quando o texto-fonte possuia a grafia preservada, foram realizadas edições de forma a modernizar a grafia atendendo aos requisitos para utilização e melhor funcionamento de ferramentas automáticas de

anotação linguística.

- Texto-Fonte com grafia modernizada: consiste de textos que já haviam sido editados e modernizados por terceiros. Neste caso não foram realizadas edições para modernização da grafia, entretanto, alguns itens foram modificados para melhor funcionamento em ferramentas de anotação automatizada por se tratarem de itens de difícil processamento computacional, esta foi chamada de Edição Técnica.
- Casos Especiais: Alguns documentos não sofreram edições dos textos-fonte para modernização de grafia, apenas a edição técnica. São casos de documentos incluídos no Corpus na primeira fase de sua construção (1998-2003), e neles a modernização completa vem sendo realizada progressivamente, privilegiando os textos que ainda não passaram pelo processo de anotação automática.

Todos os procedimentos realizados de edição e modernização dos textos seguem rigorosamente os padrões estabelecidos no Manual disponível *online* no site do Corpus.

A organização do corpus também é uma preocupação central. Os textos são categorizados e agrupados de acordo com critérios temáticos, gêneros literários, registros formais ou informais, permitindo uma organização eficiente para futuras pesquisas. A criação de metadados detalhados, como informações sobre o autor, o contexto histórico e a proveniência dos textos, também é fundamental para facilitar a consulta e o acesso aos documentos (Galves, 2018).

A construção do Corpus Histórico do Português Tycho Brahe envolve considerações metodológicas cuidadosas (Galves, 2018). A seleção de textos representativos desse período histórico é um aspecto crítico, visando garantir uma amostra diversificada e abrangente da língua portuguesa da época. A padronização da ortografia, entretanto, é um desafio enfrentado, levando em conta as variações ortográficas e as mudanças ao longo do tempo.

3 Pré-processamento

Apesar do tratamento prévio realizado no corpus pelas equipes de desenvolvimento do mesmo, é necessário um tratamento adicional tendo em vista que o objetivo é a sua utilização como entrada para o treinamento de modelos de linguagem.

Ao realizar o download do arquivo compactado contendo os textos possuía 76 arquivos, alguns dos quais possuíam duas marcações temporais. Em relação a esses arquivos com duas datas a medida adotada foi a duplicação dos arquivos, tendo cada uma marcação temporal, totalizando 82 arquivos.

As janelas temporais utilizadas para classificação dos textos seguiu a mesma classificação fornecida pelos desenvolvedores do corpus, separando por período de nascimento dos autores, em janelas de 50 anos.

A primeira etapa de pre-processamento consistiu na criação de um *dataframe* com as colunas "Período"e "Texto". A etapa seguinte realizada foi a da remoção de *stopwords* (termos frequentes e comuns que geralmente são considerados pouco relevantes para a análise textual) utilizando a biblioteca NLTK (Natural Language Toolkit) (Bird et al., 2009).

A remoção de *stopwords* é uma etapa comum de pré-processamento de texto em várias tarefas de processamento de linguagem natural pois a exclusão dessas palavras ajuda a reduzir a dimensionalidade do texto, melhorar a eficiência computacional e destacar termos mais significativos no corpus.

Outra medida tomada no tratamento dos textos foi a criação de uma função com instruções com algumas expressões regulares para a remoção de outros elementos textuais, modificação de outros, conversão de caracteres maiúsculos em caracteres minúsculos e tokenização (conversão dos textos em listas de *tokens* ou palavras), conforme pode ser observado na Figura 1.

```
text = text.lower()
text = re.sub(r'ſ', 's', str(text))
text = re.sub(r'&', 'e', str(text))
text = re.sub(r'-[ ]{2,}.*\n[ ]{2,}', '', str(text))
text = re.sub(r'[ ]{2,}.*\n[ ]{2,}', ' ', str(text))
text = re.sub(r'[àáâãäå]', 'a', str(text))
text = re.sub(r'[èéêë]', 'e', str(text))
text = re.sub(r'[ìíîï]', 'i', str(text))
text = re.sub(r'[òóôõöø]', 'o', str(text))
text = re.sub(r'[ùúûü]', 'u', str(text))

text = re.sub(r'\[.*?\]', '', str(text))
text = re.sub(r'\(.*?\)', '', str(text))

pattern = re.compile(r'\d+')
text = pattern.sub(lambda x: num2words(int(x.group()),
                                       lang='pt'), str(text))

text = re.sub(r'[^\w\' ]', '', str(text))
text = re.sub(r'\n', '', str(text))

text = [w for w in word_tokenize(text) if w not in stop_words]
```

Figura 1: Lista de regex utilizadas na função de tratamento dos textos

Além da utilização de regex para modificações

no texto, foi utilizada a biblioteca Num2Words[1] para a conversão de caracteres numéricos em texto. Essa biblioteca converte números em sua versão escrita por extenso, entretanto a mesma apresenta uma limitação em relação à língua portuguesa. A conversão é realizada utilizando apenas a numeração no masculino, sendo assim em casos onde a versão por extenso do número seria no feminino, a substituição será no masculino.

Após todas as etapas descritas acima, o *dataframe* contendo os dados pré-processados foi exportado como arquivos `txt`, onde cada linha da tabela se tornou um arquivo distinto cuja primeira linha consistia do período temporal e a segunda linha consistindo de uma lista de palavras com todas as palavras do texto, conforme Figura 2.

```
1 1350-1399
2 ['chronica', 'delrey', 'd', 'ioam', 'i', 'boa',
  'memoria', 'reys', 'portugal', 'decimo', 'primeira',
  'parte', 'contem', 'defensam', 'reyno', 'ate', 'eleito',
  'rey', 'offerecida', 'magestade', 'delrey', 'dom',
  'ioam', 'iv', 'n', 'senhor', 'miraculosa', 'memoria',
  'composta', 'fernam', 'lopez', 'anno', '_um', 'mil',
  'seiscentos', 'quarenta', 'quatro', 'lisboa', 'todas',
  'licenças', 'necessarias', 'custa', 'antonio',
  'aluaarez', 'impressor', 'delrey', 'n', 's', 'licencas',
  'p', 'or', 'mandado', 'concelho', 'geral', 's', 'offi',
  'cio', 'vi', 'priemira', 'parte', 'chronica', 'del',
  'rey', 'dom', 'ioam', 'primeiro', 'glorio', 'sa',
  'memoria', 'coposta', 'fernao', 'lopez', 'escri', 'uao',
  'puridade', 'infante', 'd', 'fernandoe', 'nao', 'tẽ',
  'cousa', 'qencotre', 's', 'fe', 'bons', 'costumes',
  'antes', 'sera', 'muy', 'poueitosa', 'nimar',
  'portugueses', 'deste', 'tẽpo', 'q', 'co', 'maior',
  'feruor', 'defendao', 'feu', 'reyno', 'imitado', 'tao',
  'gloriosos', 'ante', 'passados', 'lisboa', 'couto',
```

Figura 2: Arquivo `txt` exportado após o pré-processamento.

Os arquivos estão armazenados em pasta na nuvem e estão disponíveis para acesso via compartilhamento de link[2] ou por email.

Referências

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with python: Analyzing text with the natural language toolkit*. Beijing: O'Reilly. doi:http://my.safaribooksonline.com/9780596516499. http://www.nltk.org/book.

Galves, Charlotte. 2018. The tycho brahe corpus of historical portuguese: Methodology and results. *Linguistic Variation* 18. 49–73. doi:10.1075/lv.00004.gal.

[1] Disponível em: https://github.com/savoirfairelinux/num2words

[2] Acesso disponível através do link: clique aqui para acessar.