# 期末レポート

- 条件：コーパスを使っていること
- ～10ページ程度
- asao@lit.nagoya-u.ac.jp
- 提出期限：８月５日（水）

# 宿題

- どちらかを選んでください（両方やる必要はありません）

1. 青空文庫から、「生徒も生徒なら、教師も教師だ」や「持っていることはいるんですが」のような同語反復表現の実例をできるだけ多く見つけてください。
2. Project Gutenberg から、*The more X, the more Y.* や *The Xer, the Yer.* のような表現の実例をできるだけ多く見つけてください。

# 正規表現 (2)

# 前回の問題

- *as long as*, *as many as* など、１単語を挟んで as が2つある箇所にマッチする正規表現は？

# 後方参照

- \1 は、( ) で一度マッチした内容にもう一度マッチさせる
- (マドンナ)\1
  - 「マドンナマドンナ」にマッチ
- 括弧が複数ある時は、（開き括弧の位置の）順に \1, \2, \3, ...

# 問題

- 「くるくる」「どんどん」など、ひらがな２文字を繰り返しているところを検索するには？

# 正規表現×置換

- 各行最初の文だけを取り出したい
- 各行先頭から、最初の「。」までにマッチする正規表現は？
  - 正規表現は Greedy

# 正規表現x置換による前処理

# 問題 (1)

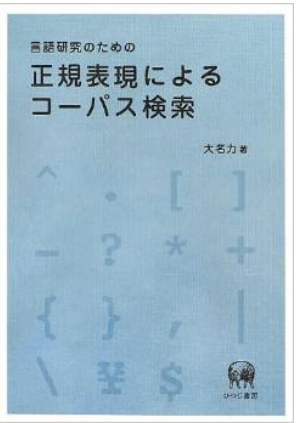
- Project Gutenberg からダウンロードした作品には文中に改行が入っている場合があります。文中の改行のみ削除するにはどうすればよいでしょうか？

## 正規表現x置換による前処理問題 (2)

- 青空文庫からダウンロードした小説には《》で囲まれた振り仮名が含まれています。振り仮名を全て削除するにはどうすればいいでしょうか？

# おすすめ

- 大名 力 (2012)『言語研究のための正規表現によるコーパス検索』
- 正規表現を説明しているウェブサイトはたくさんあるので検索してみてください




# 文字列検索の限界

- 文字列検索（正規表現を含め）では不可能なこと
  - 品詞による検索。「名作すぎる」のように「名詞＋すぎる」だけを検索、というようなことは不可能
  - 係り受けによる検索。例えば、「〜が」を含む関係節を検索、というようなことは不可能
  - より高度な前処理へ（次回）