*Perl/CHISE*による正規表現の拡張の試み

師　茂樹（花園大学）

仏教学者の悩み(1)

- 現代の“常識”が邪魔をする
- 当時の“常識”を知らない
  - 常識はテキストに残らない（暗黙知、言語該知識）

夜露死苦！

- 千年後の漢文学者はこれを何と読むだろうか？

仏教学者の妄想

- コンピュータで隠れたパターンを発掘できないだろうか
  - 「現代人には通常認知できないデータの構造性や規則性を探り出す」ことで、「「内省」(introspection)（語感）の欠如を補うことができ」る。（近藤2001）
  - 正規表現／N-gramを使って

テキストのパターン(1)

- 字面
  - 一日一善
    - `(.)[^\1]\1[^\1]`
- 意味
  - 青山白雲
  - 慈父悲母（≒慈悲父母）
- 発音（ほぼ字面）
  - アンパンマン

テキストのパターン(2)

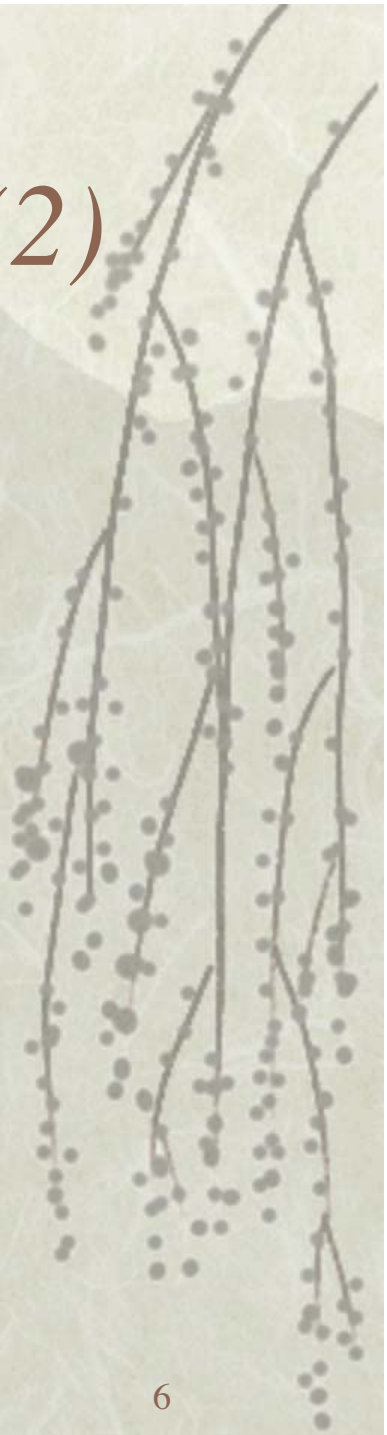
春眠不覺曉

處處聞啼鳥

夜來風雨聲

花落知多少

- 韻を踏む
- 平仄を合わせる
- etc…



仏教学者の悩み (2)

- 正規表現の機能不足
  - テキスト ＞ コード番号の集合
  - 正規表現は文字コードしか扱えない

→ 正規表現ではコード番号以外のパターンは探せない

*CHISE Project*へ

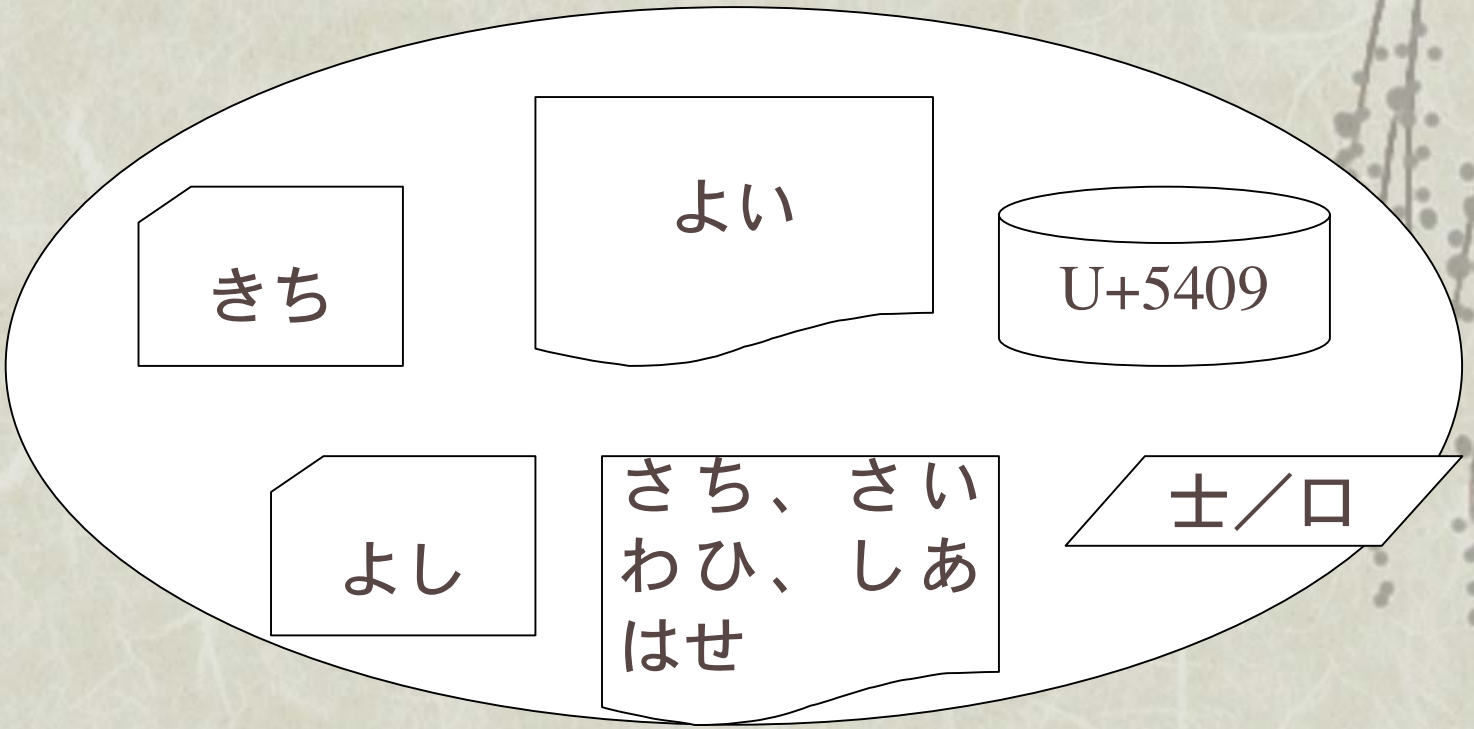
- Chaonモデル
  - 文字コード・モデルからの本質的脱却
- 様々な実装
  - XEmacs CHISE
  - Ruby/CHISE, Perl/CHISE
  - ΩTeX, KAGE
  - libchise

文字コード・モデル

- 文字の「本質」に番号をふる
  - 内包的定義
  - cf. Unicodeのcharacter

Chaonモデル

- 素性（知識）の集合による文字の定義
  - 外延的定義



Perl/CHISE

- 集合としての文字処理
- 文字コードに依存しないパターンマッチの試み

実装済みのメタ文字

- \same_*feature_n*
  - *n*番目の丸括弧の組に対応するすでにマッチした文字と、同じ素性*feature*と値の組を持つ文字にマッチ

実装方法

- 正規表現オーバーロード機能

```
sub import {
  overload::constant('qr' => \&ChiseLikeRegex);
}
  …
sub ChiseLikeRegex ($) {
  #前処理
}
```

前処理でやっていること

- \same_total-strokes_1

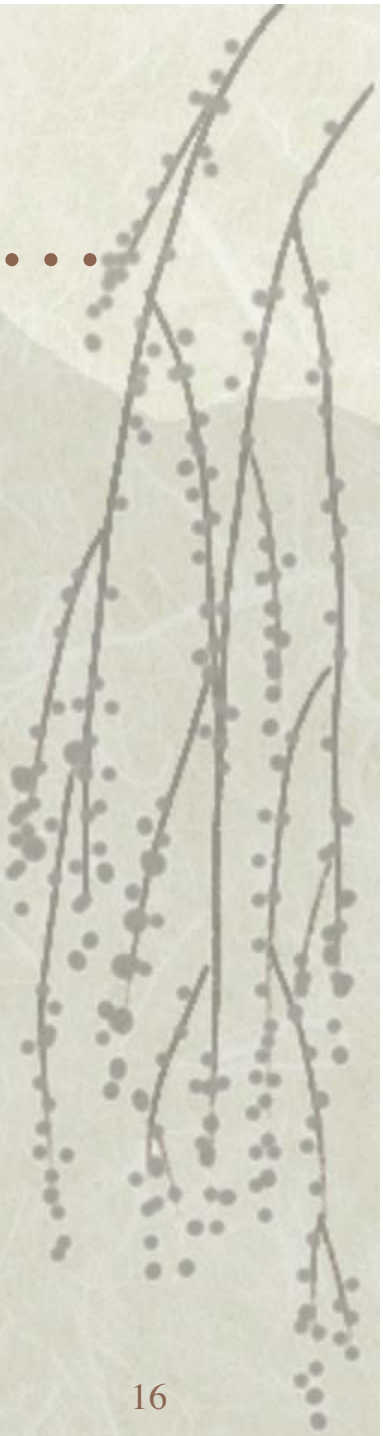
↓

```
(??{CHISE_REG::chise_backref(
$1, 'total-strokes')})
```

デモ

面壁数日...

- Overload reigns everywhere…?
  - Perl/CHISEの他のメソッド
    - 使えなくなる
  - データベース用のモジュール
    - Bus Errorが出る



I've zenned it!

- 別モジュールにする
- データベース・アクセス機能を外部化
  - libchise化すれば解決？

cf. *Hacker's Dictionary*

現在やっていること

- キャッシュ
  - あまりにも遅いので
  - libchise化するまでの一時措置？
- メタ文字の追加
  - 文字素性（とその値）を持つメタ文字
    - \p{*Unicodeのプロパティ名*}と同質

これからやりたいこと

- 素性指定の多元化
  - [\same_*f1*_1\same_*f2*_1]
    でもよいかなと思いはじめている。
- データベースの拡充
  - Kanji Database Projectも進行中
    (http://kanji-database.sourceforge.net/)

問題点

- 文字素性の文脈依存性
  - Yankees is a baseball team.
- Perlの問題
  - cf. Friedl 2003, 7.8.7.

参考文献

- Jeffrey E. F. Friedl. *Mastering Regular Expressions, 2nd Edition*. O'Reilly. 2002. 田和勝訳『詳説正規表現　第2版』（オライリー・ジャパン、2003）
- 近藤泰弘「コンピュータによる文学語学研究にできること ―古典語の「内省」を求めて―」（http://klab.ri.aoyama.ac.jp/public/paper/20010602.pdf）