CHISEプロジェクトの概要

師　茂樹・守岡　知彦

文字コード的発想の限界

- 背番号のみで管理することの限界
- 異体字テーブルの限界

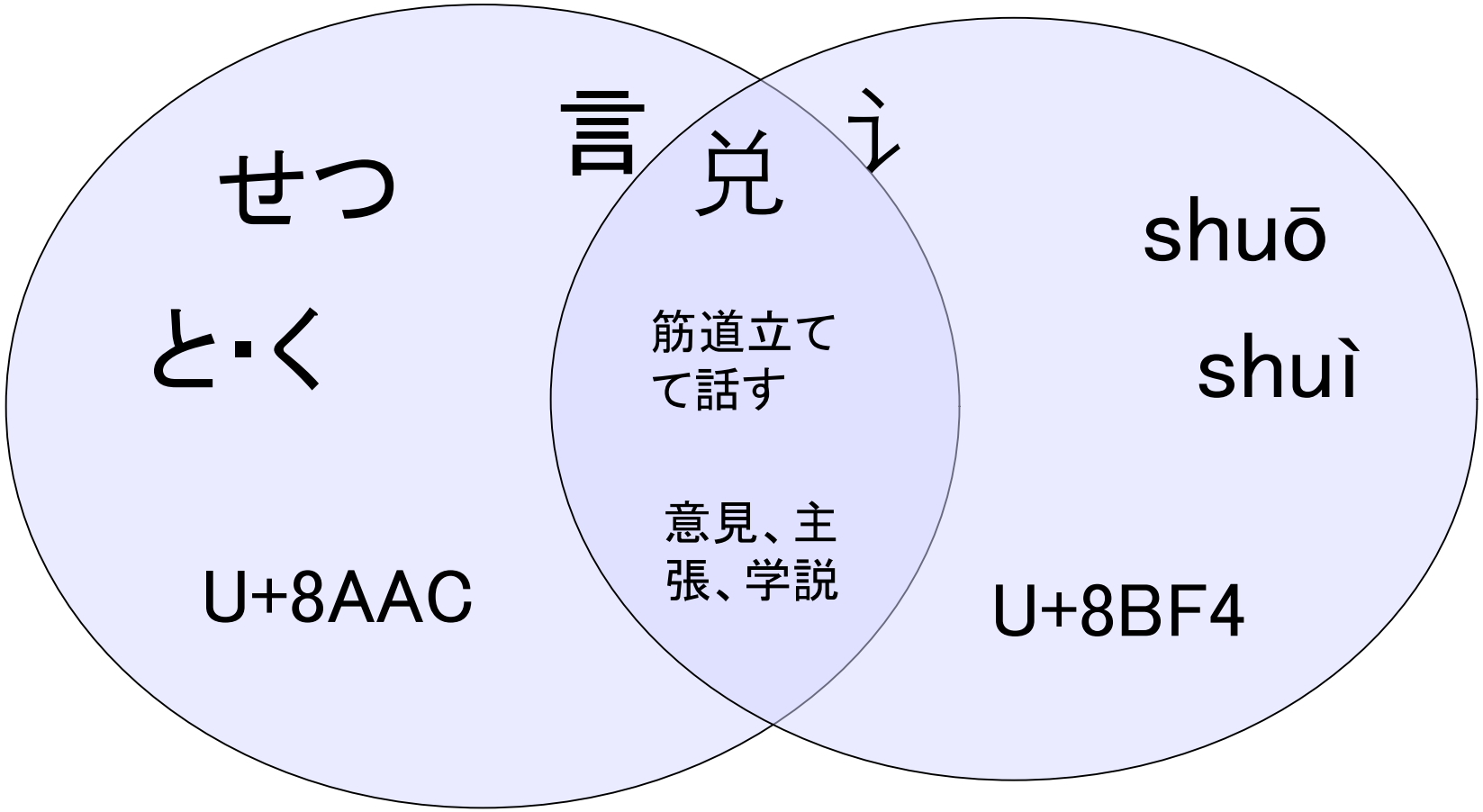
背番号のみで管理することの限界

- 「鴎」と「鷗」は同じ字なのか違う字なのか
  - 同じ字だ！
    - （かつての）JIS
  - 違う字だ！
    - TRONコード
  - 同じであるとも言えるし、ないとも言える
    - 離散集合ではない

限界の本質

- 文字をあらわす要素が1つしかない
  - 同じ／違うしか表現できない
  - かつては効率的だったが
- コンテクストから離れた「文字」論

素性の集合による文字のモデル化



素性間の関係の記述＝知識

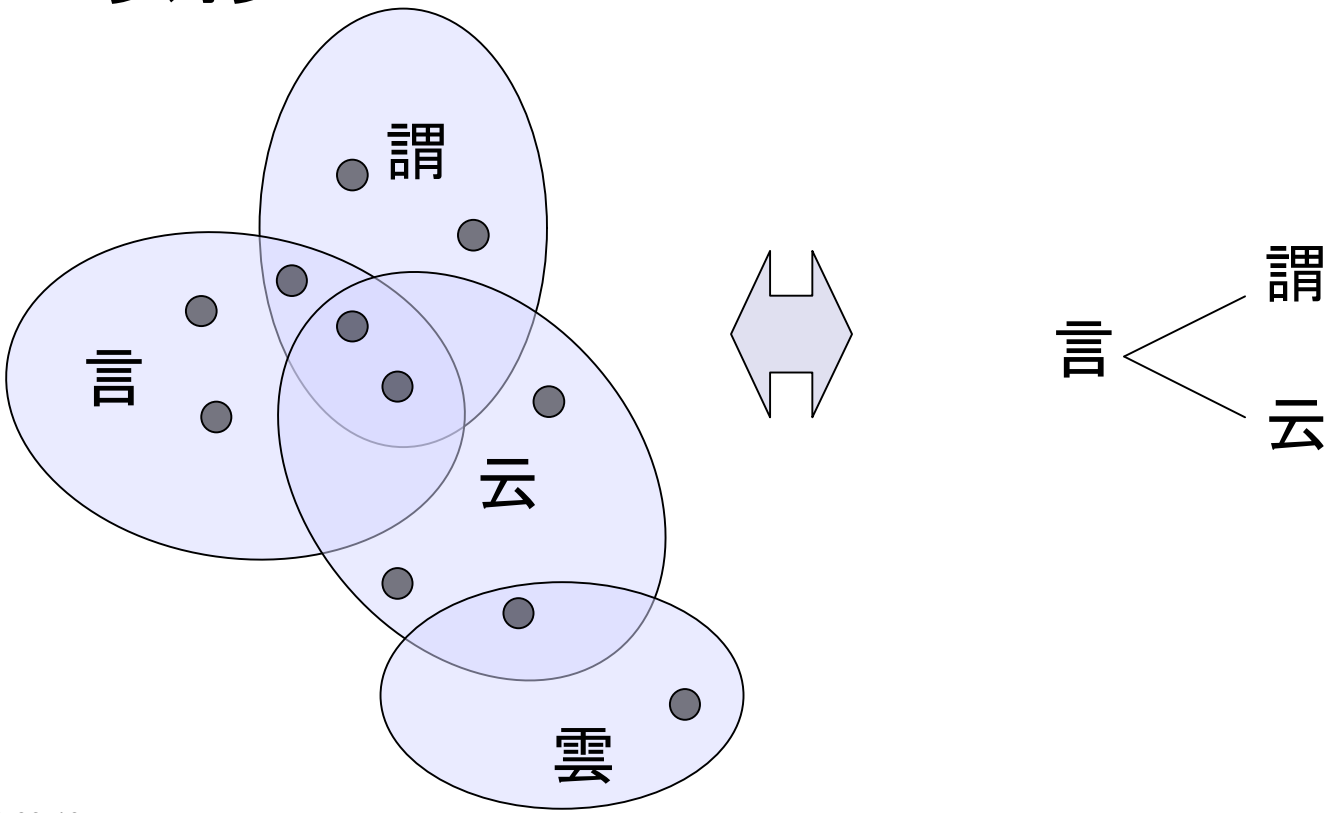
- Chaonモデル
  - 文字オントロジの記述
  - 出典（コンテクスト）が明確な知識
    - 文字コードもひとつの出典

知識とコンテクスト

- 文字のあり方はコンテクストに依存する
  - 「説」は『論語』では「悦」と同じ
- ユーザがコンテクストを自由に決定できる
  - 特定のコンテクストに縛られない

CHISE≠異体字テーブル

- CHISE
  - 多対多
- 異体字テーブル
  - 1対多のみ



CHISEの目指すもの

- 符号化文字集合モデルからの本質的な脱却
  - 文字コードの否定ではない
- “環境”の構築
  - 文書編集
    - フォント
  - 印刷・公開
  - 情報交換・知識の共有

サブ・プロジェクト (1)

- 文字知識データベースに基づく文字処理アーキテクチャの開発
  - XEmacs CHISE
  - Ruby/CHISE
  - Perl/CHISE
- 文字に関するさまざまな知識のデータベース化
  - 漢字構造情報データベース
    - CHISE-IDS 漢字検索
  - グリフ・字形情報の統合と合成
    - KAGE
  - 出典が明確な知識のデータベース化

Perlで昔やったこと

- CHISEのデータベースを使った正規表現

```
use CHISE_REG;
use utf8;
my $target = '山川';
if ($target =~ /(.)\same_strokes_1/) {
  print "matched!\n";
} else {
  print "unmatched...\n";
}
```

サブ・プロジェクト (2)

- 文字知識情報の数理的解析と可視化
  - CHISE漢字連環図
- 文字データベースと連携した組版システム
  - Ω/CHISE

参加者募集！

- http://www.kanji.zinbun.kyoto-u.ac.jp/projects/chise/