Unicodeに見る
文字コード国際化の現状と課題

政策情報学部

0940125
氏名　李　蓮

結論

Unicodeが持つ問題点について解決策を検討

■解決案

民族のアイデンティティ＝言語や文字

使い続けたい・続けさせてあげたい

情報化

文字コードとして残す

Unicodeに入れてもらう

背景：インターネットの普及

- 英語圏の人々にとって完璧な仕組み

- 漢字を使う日本、中国、台湾、韓国など不便
  - ISO 2022 JP: 日本語を電子メールで使う方法の標準化
  - 各国の文字のエンコードを個別に考える必要があった

背景「満族」

「満族」とは

—中国の55の少数民族の一つ

—1000万人を超える人口

—その言語や文字ほとんど失われる

目的

文字コードの国際化の改善を検討することを通じて、民族のアイデンティティを維持し、言語や文字を使い続けさせること

今まで調べたこと

- 文字コードの今までの歴史
  - 主に東アジア
- 民族と文字・言語の関連性
  - 満族の場合
  - パキスタンの状況

現状：EUCにおける東アジアの文字

- EUCは、各国の文字をすべてサポートしている
- 東アジアにおける主要文字と言語
  - 日本（EUC-JP　JIS拡張漢字には11,233字が収録）
  - 中国（EUC-CN　GB18030-2005には70,244字が収録）
  - 台湾（EUC-TW　CNS 11643-1992には48,711字が収録）
  - 韓国（EUC-KR　KS X 1001-2002には8,227字が収録）

文字コードについて

- EUC
  - 複数のバイト幅を用いて漢字のような膨大な数の文字でも扱うことができる
  - EUCは正しい実装
- Unicode
  - 全世界の文字を16ビットで表現することを目指す規格
  - Unicodeは問題あり

Unicode問題点

- 漢字統合
  - 字形的に同一もしくは些細な字形差の文字同じコードポイントに割り振る
- CJK漢字統合の影響
  - 文字の区別ができない
  - 民族のアイデンティティが守れない

文字の区別ができないことによる影響

- お父さんは「李　春燮」から「李　春変」に
- 「消えた5000万件の年金記録」
  - 漢字仮名変換システムによる記録の誤り

日本の漢字と韓国の漢字

日本の漢字

- 日本で生まれた漢字　:
- 中国から伝わった漢字：宅配便など

韓国の漢字

- 主に必要に応じ名前などに漢字が使われる

  例えば김 상미（金　徜美）

- ニュースなどで外国の名前に漢字が使われる

  例えば 日 함박눈이 내려

  (日本で大雪が降った）

台湾の漢字と中国の漢字

台湾の漢字

繁体字

例えば我是學生（私は学生です）

中国の漢字

簡体字

例えば我是学生（私は学生です）

漢字統合の理由

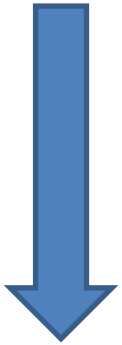
なぜ小さくまとめたがる？

考えられること

- 欧米人は状況をわかってない
- コード化→効率化→なるべく少ない情報量にまとめる方向

CJK漢字統合 の解決策に向けての日本の動き

「Unicode IVS　Add-in for Microsoft Office」の公開



58000の異体字を含む表示・印刷・編集が可能に

解決策の提案

ー空間を広げて統合させている字を全部バラバラにする

ーUnidodeに入っていない文字は入れる

民族のアイデンティティが保たれる

まとめ

- Unicodeが持つ問題点について解決策を検討
- 過去の民族と言語・文字の状況についての調査

今後の課題

「Unicode IVS　Add-in for Microsoft Office」公開以降の研究者の動き

中国の少数民族たちが使っている22種類の文字は未調査

独立した民族でありながら、自らが持つ言語と文字を使用できていない状況の民族を調査

ご清聴ありがとうございました。