第 12 回：まとめ

1. 授業のまとめ

1.1. 外国語「処理」とは

基本：「読む・書く・保存する・印刷する」

入力システムの利用法
エンコード方式の理解
フォントの利用法
さまざまな文字の入力方法

⇒ 論文・レポートや，資料となるデータの作成
⇒ 電子メール，ウェブページ等，メディアや用途に応じた外国語文書作成

応用：「加工する」

⇒ ファイル形式の使い分け (Word 形式，RTF 形式，テキスト形式など)
⇒ エンコード方式の変換 (ローカルなエンコード方式⇔Unicode；中国語簡体字⇔繁体字)
⇒ Unicode 変換方式の使い分け (UTF-8, UTF-16)
⇒ 文字の置き換え (e.g. 文字参照の利用： **ä** ⇒ **ä** →第 9 回資料を参照; **ä** を **a% や a#** で表現する，など) [1]
⇒ ツールを使ったテキスト処理 (例：ソート→第 11 回資料を参照)

※ ローカルなエンコード方式間の変換

- 中国語など，漢字圏の言語のテキストを相互に変換する場合に重要。大学の PC には「RTF コンバータ」というエンコード方式の相互変換用のツールがインストールされているので, 利用してみよう ([スタート]→[プログラム]→[Harigaya's Converter]→[RTF コンバータ]) [2]。

1.2 この授業で取り上げなかったこと

- テキスト処理用プログラミング言語 (grep, sed, awk, perl など) の利用
- 各言語の論文の一般的な書式や電子メールの作法
- PDF (Portable Document Format)： Adobe の開発した多機能な電子文書形式。

---

[1] 特殊な文字は，例えば ASCII の基本英数字・記号を使って置き換えることでローカルなエンコード方式でも利用できるようになる。パソコンで文字を一時的に置き換えて保存する場合は確実に元のデータに戻せることが重要。例えば，ドイツ語のウムラウト ä, ö, ü を ae, oe, ue と置き換えた場合，置換機能を使ってウムラウトに戻そうとすると，本来 ae でつづられていたものまで ä に変換されてしまう。そこで，例えば文字参照 (第 9 回資料§1.3.を参照) を利用したり，a#, a% といった通常使わない文字列でそのような文字を入力しておけば，安全にウムラウトを置換で再現できる。

[2] 第 4 回授業で紹介した Word の「中国語の翻訳」ツールを利用してもよいだろう ([ツール]→[その他の構成ツール]→[中国語の翻訳])。単純に EmEditor などを使って Big5 のテキストを GB2312 の文字エンコードで保存しなおすだけでは漢字は正しく変換されない (文字の問題については Unicode の漢字の統合の問題 (第 10 回資料) も参照)。RTF コンバーター (針谷壮一先生作成) は簡体字⇔繁体字の漢字の変換を非常に上手におこなってくれる，Kadai サーバの[schiba]→[2004fl]→[No12]フォルダに，全く同じ内容の中国語の単語リストを，簡体字中国語 (GB2312) と繁体字中国語 (Big5) で入力したサンプルファイルがある (それぞれ simplifiedchinese.txt と traditionalchinese.txt というファイル名が付いている)。RTF コンバータを利用して一方の文字エンコード方式から他方に変換し，変換した結果を比べてみよう。もちろん，RTF コンバータの変換結果が間違っていたり, 自分の意図した漢字に変換されていない可能性もあるので，最終的には変換結果をよく確かめることが重要である。

閲覧用の OS を選ばないほか, 多言語テキストにも対応。PDF の閲覧ソフト Adobe Reader は多くの基本ソフトに対応しており, ダウンロードして無償で利用できる。

- (日本語における) 漢字の問題： 日本工業規格 JIS が定める漢字には，Shift JIS で扱える JIS 第 1 水準漢字 (2965 字)，第 2 水準漢字 (3390 字) に加え，JIS 第 3 水準漢字 (1908 字)，第 4 水準漢字 (2436 字) や補助漢字 (漢字 5801 字 + 漢字以外の文字 266 文字) があり，人名や地名などに使われる漢字などが含まれている。JIS 第 3/4 水準漢字や補助漢字は Unicode では利用可能だが，Shift JIS では利用できない。)
- 日本語以外の言語用に開発されたソフトウエアの具体的な利用方法 (英語用以外のソフトには Windows のシステムロケールの変更が必要なものもある→第 2 回資料参照)

1.3. パソコンでの外国語テキスト利用のポイント

専攻言語の勉強を進めていくと，レポートやプレゼンテーション，論文などはもちろんだが, 外国語のテキストをデータとしてパソコンに蓄積したいことがある (例えば, 外国語の用例集や用語集を作ったり，文献一覧を作ったり，印刷された文献やメモを入力して整理したり…)。資料として使うために作成するデータは, どうやって作成したらよいだろうか。

- 利用するアプリケーションと, ファイルの形式をまず考えよう (利用するアプリケーションは Word? Excel? EmEditor? 保存形式は Word 形式? Excel 形式? RTF (第 5 回資料参照)? それともテキスト形式? テキスト形式ならエンコードは Unicode (UTF-16, UTF-8)? それともその言語・地域のローカルなエンコード方式? など)。
  - データを複数のアプリケーションで利用する場合や，電子メールや Web ページなど，異なるメディアで利用する場合には，それらのソフトで共通して利用できるファイル形式やエンコード方式を選択することが重要。
- テキスト中に複数の言語を混在させたり，特別な文字・記号を使ったりする必要があるか，よく考えること。
  - **一旦 Unicode を使って多言語を混在させたり，その言語のエンコード方式に含まれていない記号を使うと，そのテキストはローカルなエンコード方式を使ったテキストとしてそのままでは保存できない。**
  - 2 ヶ国語の語彙集のように, 複数言語を入力する場合は Excel のセルを利用して入力場所を分離するなど，入力方法を工夫するとよい。
- 作業前のオリジナルデータをバックアップする (エンコードの設定に失敗して, 復元できない文字化けが起こってから後悔しても遅い！)。

1 クラスの期末試験について：詳細は教務から出される期末試験の掲示を参照すること

期日： 学期末試験期間中，60 **分** (7 **月** 27 **日** (**火**) 11:40〜12:40 (3 **時限目**) **の予定**)
場所： 1501 **の予定**
持ち込み： 可
問題： 筆記 (**パソコンは使用しない**)**。主に以下のテーマから出題する予定。**

- Windows XP **の多言語機能について**
- **エンコード方式の種類** (ASCII, **各言語･地域のエンコード方式**, Unicode)
- **外国語テキスト文書を扱う際に注意すること**
- **さまざまなファイル形式** (**バイナリ形式** vs. **テキスト形式，**Word **文書とテキスト文書**, HTML **文書など**)
- Unicode **について** (Unicode **とは，「**Unicode **アプリケーション」とは，**Unicode **の変換方式，**Unicode **を利用する際の注意点**)