第 13 回：文字コードの変換

本日のポイント：

- 文字エンコード方式間の変換
  - ローカルな文字エンコード方式から Unicode へ
  - Unicode から，ローカルな文字エンコード方式へ
  - Unicode 間の変換形式の変更

1. 文字エンコード方式間の変換

1) ローカルな文字エンコード方式から Unicode へ

2) Unicode から，ローカルな文字エンコード方式へ

- ローカルなエンコード方式に収録されていない文字がある場合，どうするか？

  解決例：HTML の数値文字参照を用いる [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 11 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，置換によってウムラウトを確実に再現できる。

  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，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ことに注意。文字の問題については Unicode の漢字の統合の問題（第 6 回資料）も参照すること。

- 大学の PC には「RTF コンバータ」(針谷壮一先生作成）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（[スタート]→[プログラム]→[Harigaya's Converter]→[RTF コンバータ])。

- 簡体字⇔繁体字中国語を変換する場合は，第 4 回授業で紹介した Word の「中国語の翻訳」ツールを利用してみてもよいだろう（[ツール]メニュー→[その他の構成ツール]→[中国語の翻訳])。

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。

3) Unicode の中で, 変換形式を変更する（UTF-16 ⇔ UTF-8, UTF-16LE ⇔ UTF-16BE など，UTF-16 の場合には BOM の付加が必須であることに注意)

[1] 後で紹介する RTF コンバータには，変換できない文字を数値文字参照で置き換えてくれるオプションがある。

担当：千葉庄寿（schiba@reitaku-u.ac.jp）

実習１： Kadai サーバの[schiba]→[2006fl]→[No13]フォルダに，フィンランド語と日本語のバイリンガルテキスト finnish.txt が UTF-16LE で保存されている。

1) finnish.txt を EmEditor で開き，文字エンコード方式を Shift JIS に変更し，finnish_sjis.txt という名前をつけて「強制的に」保存しなさい。
2) finnish_sjis.txt を一旦閉じ，再度開いて ä がどのように保存されているかを確認しなさい。
3) finnish.txt を EmEditor で開き，文字エンコード方式を ISO-8859-1 に変更し，finnish_latin.txt という名前をつけて「強制的に」保存しなさい。
4) finnish_latin.txt を一旦閉じ，再度開いて日本語がどのように保存されているかを確認しなさい。
5) ä を ae に置換し，finnish_sjis2.txt という名前をつけて Shift JIS で保存しなさい。
6) finnish_sjis2.txt の ae を ä に再度置換しなおし，finnish2.txt という名前をつけて UTF-16LE で保存しなさい。
7) finnish.txt と finnish2.txt を比較し，ä → ae という置換で発生する問題を確認しなさい。finnish2.txt は一旦閉じなさい。
8) RTF コンバータを用い，finnish.txt の ä を文字実体参照を使って置き換えるよう設定し，finnish_sjis3.txt という名前をつけて Shift JIS で保存しなさい。
9) RTF コンバータを用い，finnish_sjis3.txt の文字実体参照を元の戻すように設定し，finnish3.txt という名前をつけて UTF-16LE で保存しなさい。

実習２： RTF コンバータは簡体字⇔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]→[2006fl]→[No13]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。また，2 種類のテキストを Word に貼り付けて，Word の「中国語の翻訳」ツールを使った変換も試してみよう。

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 31 日 (月) 13:30〜14:30 (4 時限目) の予定)
場所： 1401 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
- 文字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)
- 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)
- 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) および各ファイル形式で外国語テキスト文書を扱う際に注意すること
- 基本的な正規表現について (←次週扱います)

以上