第 1 回：ガイダンス

講義内容

- 「テキスト (text, textual data)」：コンピュータで扱うことのできるデータの中で，単純かつ最も汎用的な形式
- テキストデータを扱うための基礎知識の習得
  1. テキストデータを自在に加工・整形するために役立つ知識や技術
  2. テキストデータの特性と可能性の考察。
  3. さまざまなソフトウエアにおけるテキストデータの取り扱い
     - エディタや表計算ソフト
     - テキスト処理用の各種ツール
     - プログラミングによる柔軟なテキスト処理
- 「共通の知識基盤」としての情報処理知識
  - ✧ 専門分野としての「自然言語処理 (NLP, natural language processing)」
  - ✧ 人文系の研究者と情報工学・自然言語処理の専門家との共同研究の可能性
- 自分の研究テーマを見据え，授業内容を自身の研究におけるデータの作成や分析にどのように応用できるかを考えながら，貪欲に知識を吸収していくことを希望します。
- 統計処理は，頻度を数える，といったごく基本的なもの以外扱いません。

授業計画

回	日付	授業内容
1	04-18	ガイダンス
2	04-25	テキストエディタの利用
3	05-02	テキストデータと文書構造の表現：固定長テキスト，タブ区切りテキスト，CSV, HTML, XML
4	05-09	正規表現による検索･置換 (1)
5	05-16	正規表現による検索･置換 (2)
6	05-23	文字エンコード方式と Unicode の基本
7	05-30	フォントの利用とその問題：コードとフォントの関係，外字とは，Unicode とのかかわり
8	06-06	表計算ソフトを用いたテキストデータの処理 (1)
9	06-13	表計算ソフトを用いたテキストデータの処理 (2)
10	06-20	データベースの作成と利用
11	06-27	プログラミング言語 Perl によるテキスト処理 (1)
12	07-04	プログラミング言語 Perl によるテキスト処理 (2)
13	07-11	プログラミング言語 Perl によるテキスト処理 (3)
(レポート提出)		

評価方法

- 出席，課題，レポート (テーマは未定) で総合的に評価。

教科書

- プリントその他を配布する。
- 以下の書籍は(多少古いが)パソコンを活用して研究を進めるうえで参考になるのでできるだけ読んでおくことをすすめる：

中尾浩著 (1996) 『文科系のパソコン技術』中公新書

- 以下は文字コード一般に関する比較的まとまった情報が得られるので参考文献としてすすめる：

  清水哲郎著 (2001) 『図解で分かる文字コードのすべて』日本実業出版社

参考文献 (和書を中心に)

※ 最新刊や受講者のテーマに関連した書籍，有益な URL などは授業で随時紹介する。

言語・文学研究とパソコン：

伊藤雅光 (2002) 『計量言語学入門』大修館書店
上田博人 (1998) 『パソコンによる外国語研究への招待』くろしお出版 (品切れとのこと)
漢字文献情報処理研究会編 (2000) 『電脳国文学』好文出版
漢字文献情報処理研究会編 (2001) 『電脳中国学 II』好文出版
漢字文献情報処理研究会編 (2000-2004) 『漢字文献情報処理研究』1-5. 好文出版
※ 漢字文献情報処理研究会 URL: http://www.jaet.gr.jp
『日本語学』2003 年 4 月臨時増刊　特集「コーパス言語学」明治書院
Hunston, Susan (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

テキスト処理と正規表現：

大名力「正規表現によるテキスト検索」オンライン資料 URL:
http://infosys.gsid.nagoya-u.ac.jp/~ohna/re/index.html
齊藤俊雄 他 (1998, 2005²) 『英語コーパス言語学：基礎と実践』研究社
中尾浩 他 (2002) 『コーパス言語学の技法 I テキスト処理入門』夏目書房
同 (2004) 『コーパス言語学の技法 II 言語データの収集とコーパスの構築』夏目書房
Friedl, Jeffrey E. F. (歌代和正監訳, 2003) 『詳説正規表現』第 2 版 オライリー・ジャパン

テキストによる文書構造の表現方法としての HTML (HyperText Markup Language), XML (Extensible Markup Language)**：**

神崎正英 (2000) 『ユニバーサル HTML/XHTML』毎日コミュニケーションズ

プログラミング言語 Perl (**テキスト処理に優れる**) **の入門書：たくさんあります**

大名力「Perl によるテキスト処理入門」オンライン資料 URL:
http://infosys.gsid.nagoya-u.ac.jp/~ohna/perl_lesson/index.html
クロス, デイビット (宮川訳, 2003)『Perl データマンジング：データ加工のテクニック集』ピアソン・エデュケーション
武藤健志・トップスタジオ (2004 年) 『独習 Perl』第 2 版. 翔泳社
平田 豊 (2004) 『Perl トレーニングブック―これから始める人の Perl プログラミング練習帳―』ソーテック社
Christiansen, Tom 他 (Shibuya Perl Mongers 監訳, 2004) 『Perl クックブック』第 2 版. (全 2 巻) オライリー・ジャパン

多言語処理・文字コード：

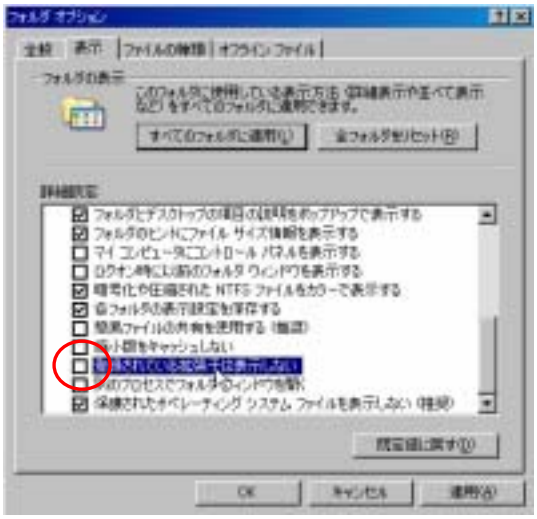
清水哲郎著 (2001) 『図解で分かる文字コードのすべて』日本実業出版社
加藤弘一 (2002) 『図解雑学文字コード』ナツメ社
Ken Lunde (小松・逆井訳, 2002) 『CJKV 日中韓越情報処理』オライリー・ジャパン

文字コード字典

芝野耕司 (2002) 『JIS 漢字字典』増補改訂版. 日本規格協会
ユニコード漢字情報辞典編集委員会編 (2000) 『ユニコード漢字情報辞典』三省堂

担当：千葉庄寿 (CHIBA Shoju)　　e-mail: schiba@reitaku-u.ac.jp

準備体操：ファイル名と拡張子



Windows では，ファイルの種類をファイル名の末尾の記号で表す。この記号を**拡張子** file extension という。Windows は拡張子によってどのアプリケーションで開くかをあらかじめ登録し，起動しやすくしている。この対応を**関連付け** association といい，関連付けられたファイルは，「マイコンピュータ」からファイルをダブルクリックするだけで開くことができる。

大学 PC では，ログオン直後の状態では，関連付けられたファイルの**拡張子は省略されて表示される**。拡張子を含んだ「正式な」ファイル名を知らないために，しばしばトラブルが起こる。拡張子を含むファイル名を表示する方法を覚えておこう。

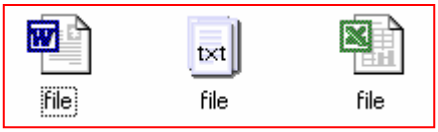
1. 「マイコンピュータ」を開き，メニューバーの[ツール]から[フォルダオプション]を開く。
2. 「表示」タブを選択する。
3. **「登録されている拡張子は表示しない」という項目のチェックをはずす**（右上図を参照）。
4. 「ＯＫ」ボタンを押す。これで拡張子付きのファイル名が表示されるようになる。

※**この設定は，大学のパソコンではログオンする度におこなわなければならない。**
自分のパソコンの場合，一度設定すれば同じ設定が適用される。

> 重要：この授業では，ファイル名を指示する場合は，原則として**拡張子を含んだファイル名**を使います。

実習：「マイコンピュータ」を開き，file_server **の** Kadai **の** schiba **フォルダにあるフォルダ「**kenkyuu2005**」を** file_server **の** Home **にコピーしなさい** (kenkyuu2005 **の中には** No1 **というフォルダがあり，ファイルが** 3 **つ入っているはず**)。

「マイコンピュータ」から「No1**」フォルダを開くと，中に入っている** 3 **つのファイルは全て** file **という名前で表示されている。上の拡張子の説明を読み，拡張子と，拡張子を含めた「正式な」ファイル名を調べよう。**



拡張子なしファイル名	ファイルの種類	拡張子	拡張子つきファイル名
file	Microsoft Word 文書		
file	Microsoft Excel ワークシート		
file	テキストドキュメント		

※自宅のパソコンと大学のパソコンでは，表示される「ファイルの種類」の名称やファイルアイコン，関連付けられているソフトウエアが異なる場合がある (例えば，大学のパソコンでは，「テキストドキュメント」は EmEditor に関連付けられている)。