第 12 回: Excel を用いた多言語データの作成･編集

本日のポイント:

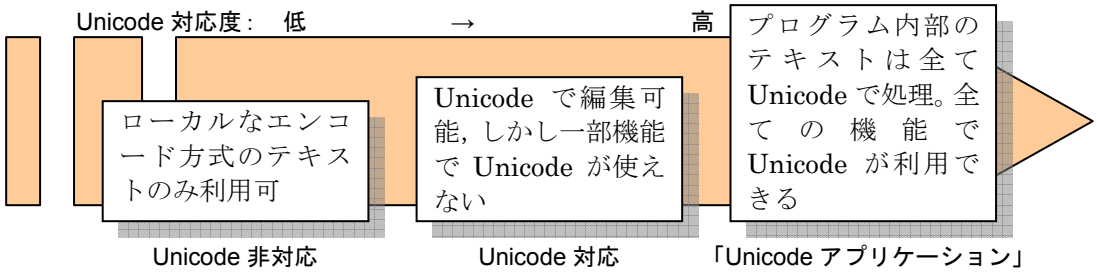
- 「Unicode アプリケーション」としての Excel
- Excel の復習
  - データの入力
  - データの並べ替え（ソート）
  - データの絞り込み（オートフィルタ）
- 多言語データの加工例：Excel と Word のソート機能
- 補足：Unicode を用いたテキストファイルの作成・保存と編集―Word の場合

0. 「Unicode アプリケーション」としての Excel

第 5 回，第 6 回で学んだように，Unicode は近年急速に普及している新しい文字集合であるが，常に利用できるとは限らない。Unicode 文字を入力・表示するためには，**OS のほか，「フォント」「入力システム」「アプリケーションソフトウエア」の全てが Unicode に対応している必要がある**。Windows XP は Unicode に対応しており，多くの国・地域の言語に対応した入力システムと Unicode フォントが標準で備わっているが，アプリケーションソフトウエアの Unicode 対応状況はさまざまである。

これまで授業で利用した Word 2003，EmEditor，また Windows XP に付属する「メモ帳」，「ペイント」[1]，Internet Explorer などは，Unicode テキストを処理できるばかりでなく，プログラム自体が Unicode で書かれており，Unicode にほぼ完全に対応した**「Unicode アプリケーション」**といえる。

今回扱う Excel 2003 も「Unicode アプリケーション」であり，日本語はもちろん，さまざまな言語のテキストを Unicode で扱うことができる。



「Unicode アプリケーション」である，ということは，Unicode のいろいろな文字をそのソフトの中で使うことが可能である，ということであって，言語データに必要な処理を何でも行ってくれる魔法のツールである，ということではない。今回は，Excel と Word の文字の並べ替え(ソート)機能を自分のおこないたい作業に基づいて比較し，使い分けてみよう。

[1]「ペイント」の文字入力機能で多言語テキストを利用することができる。描画ソフトは Unicode 対応が遅れており，例えば Adobe 社のプロむけドロー系描画ソフト PhotoShop は新バージョン(CS) でやっと Unicode に対応した。

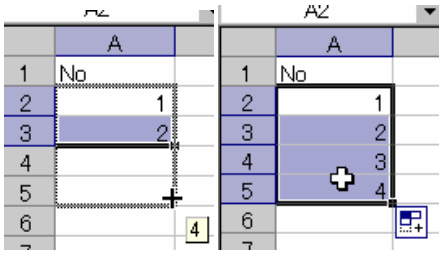
担当：千葉庄寿（schiba@reitaku-u.ac.jp）

1. Excel の復習

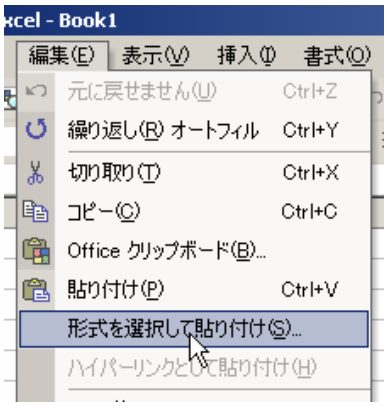
一定の規則でデータをたくさん集めたものを**データベース**という。データベースの規模はデータの件数やデータの内容，構造により大きく異なる。数 10～数千件のデータ程度で構造が比較的シンプルな小規模のデータベースは，表計算ソフト Excel を用いて簡単に作成することができる。

1.1. データの入力

- Excel のワークシートを使ってデータベースを作成する場合，縦方向(「**行**」)をデータの単位とし，項目を横方向（「**列**」）に入力するのが普通である。
- データベースにはタイトル行を作成するよう習慣づける。後で何のデータか分からなくなってしまうことがある。
- リストの入力のヒント
  - 列方向にセルを移動： Tab キー
  - 行方向にセルを移動： Enter キー
  - 再編集： セルをダブルクリックするか，F2 キーを押す。
- いくつかの便利な機能
  - **オートコンプリート**：すでに入力されているデータを候補として提示し，入力の手間を省く（Enter を押して自動入力，BackSpace で候補をキャンセルし，通常入力を続行)。また Alt + ↓ キーで候補の一覧を表示できる。
    ※ 解除は［ツール］→［オプション］の「編集」タブからおこなう。
  - **オートフィル**：各フィールドに設定された書式や計算式は，データを追加した際に隣接したセルに自動的にコピーされる。連番（1, 2, 3…）のような特定のパターンで数字を入力する場合には，複数のセルにパターンを入力し，それらのセルをマウスで選択しておいてオートフィルをおこなう(右図)。

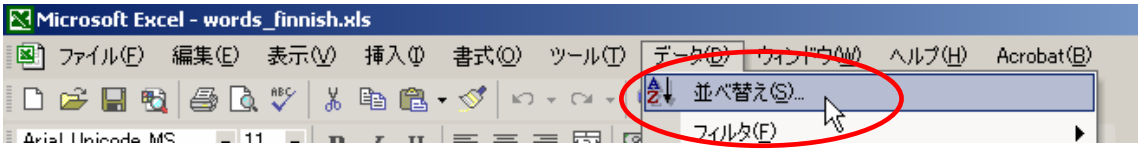


  - **フォーム**：［データ］→［フォーム］… 入力のほか，簡単な検索に使える。
  - **高度なコピー「形式を選択して貼り付け」**：テキストだけ，データの書式だけ，計算式ではなく計算した結果だけ，など，コピーしたい内容はさまざま。［編集］→［形式を選択して貼り付け］で詳細に貼り付ける内容を選べる（右図)。



1.2. データの並べ替え（ソート *sort*）

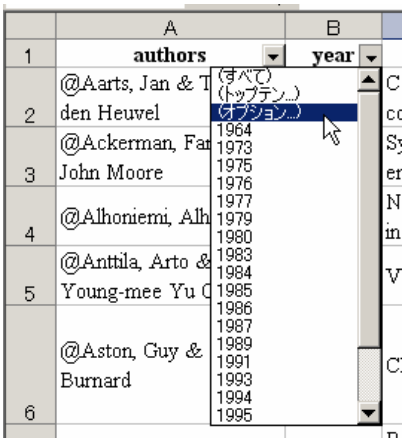
Excel は数値の大小や日付，アルファベット順などでデータを並べ替える(ソートする)機能をもつ。



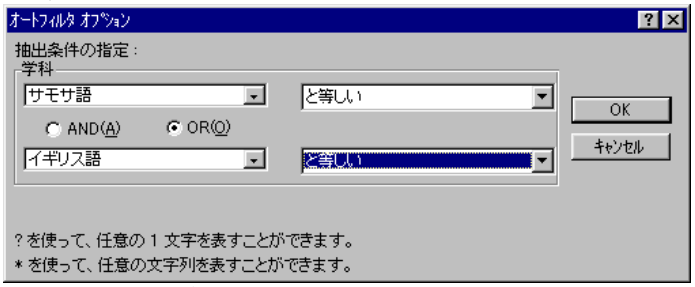
- Excel は「昇順 ascending」「降順 descending」というふたつの方法でソートできる。
  ⇒ 昇順・・・数字の小さい順，文字ならば abc 順・50 音順にデータを並べ替える。
  ⇒ 降順・・・数字の大きい順，文字ならば abc 順・50 音順の**逆順**（c，b，a，お，え，う，い，あ）にデータを並べ替える（もっとも，文字の降順はあまり使うことはない [2]）。
- ソートの方法
  (1) 並べ替えの基準となる列（基準列）を決める。
  (2) ソートする基準列（列のどの部分でもよい）をクリックする。（特定の列だけを選択しないこと！その列の中だけでソートがおこなわれてしまう！）
  (3) 以下のボタンをクリックする。　昇順：  　降順： 
  (4) データが並べ替えられる。
- 複数の列で並べ替えるなど，より細かい指示を行いたい場合には［データ］→［並べ替え］を選択し，ソートのダイアログボックスを表示する。
- 【注意】ソートを一旦おこなうと，元の順番を復元できなくなる恐れがある。入力した順番に戻す必要がある場合を考え，通し番号や入力日時など，一貫した情報をつける列を用意しておくとよい。

1.3. 絞込み検索（オートフィルタ）※今回の授業では扱わないので，各自練習しておくとよい。

Excel の一覧表にあるデータのなかから，特定の条件を満たすものだけを取り出して表にするには，Excel の「オートフィルタ」という機能を使用する。



- [データ]→[フィルタ]→[オートフィルタ]と選択してオートフィルタを起動する。
  ➢ 【注意】オートフィルタを利用する場合には，タイトル行の各列に必ず見出しをつけておくこと。
- 各見出しに絞り込み用のプルダウンメニューがつく。
- 細かい条件を設定する場合には，プルダウンメニューから「オプション」を選択する。「オプション」では 2 つまでの条件を AND, OR 検索できる。



「抽出条件」には Unicode の多言語テキストを利用することができる。

- 複数のフィールドに抽出条件を指定して複雑な絞込み検索をおこなうことができる。不要な条件を解除する場合には，プルダウンメニューから「すべて」を選択する。
- オートフィルタの解除：オートフィルタの機能を使い終わったときは，メニューから再度，［データ］→［フィルタ］→［オートフィルタ］と選択する。

[2] 逆順は，「逆引き」とは異なる。逆引きは，単語の後ろの文字から昇順にソートしたもの。

担当：千葉庄寿（schiba@reitaku-u.ac.jp）

2. 多言語データの加工例：Excel と Word のソート機能

Excel は表計算ソフトであるので，数値の大小や日付，アルファベット順など，基本的なソートの機能も持っている。さらに，Excel はデータベースの検索機能（§ 1.3.）をもち，規則的なデータの入力を効率よくおこなうことができる（§ 1.2.)。

Excel2003 は Unicode 対応アプリケーションであり，日本語はもちろん，さまざまな言語のテキストを Unicode で扱うことができる。しかし，Excel がもっているソート機能は，基本的に文字のコード値の大小を利用するものなので，各言語固有の文字の順番に対応したきめ細かなソートはできない。

- アルファベット言語のテキスト:辞書の見出し語の順番は,言語によって大きく異なる。以下のヨーロッパ各言語のアルファベット順を比べてみよう。

スペイン語	A, B, C, Ch, D, E, F, G, H, I, J, K, L, Ll, M, N, Ñ, O, P, Q, R, S, T, U, V, W, X, Y, Z
ドイツ語	A (a, ä), B, C, D, E, F, G, H, I, J, K, L, M, N, O (o, ö), P, Q, R, S (ss, ß), T, U (u, ü), V, W, X, Y, Z
フィンランド語	A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V (v, w), X, Y, Z, Ä, Ö, Å
フランス語	A (a, à, â), B, C (c, ç), D, E (e, é, è, ê), F, G, H, I (i, î), J, K, L, M, N, O (o, ô), P, Q, R, S, T, U (u, û), V, W, X, Y, Z

Excel はこのような言語ごとのアルファベット順でのソートには対応していない。

- 漢字圏の言語のテキストのソート：漢字圏の言語の場合，「ふりがな」や中国語のピンインなど，一貫した並べ替えをおこなうための項目をあらかじめ用意しておくことで，Excel でも並べ替えができる。

そこで,アルファベット言語のテキストについて,Excel で作成したデータを Word2003 を使いソートしてみることにする。Word2003 (および 2000, 2002) には，単純な文字コード順ではなく，ソートに使用する言語を指定できるより高度な機能が搭載されている。

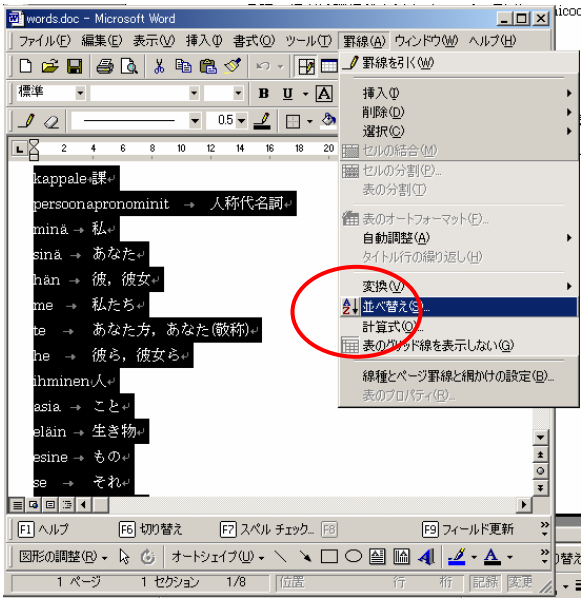
1. Excel のワークシートを「Unicode テキスト」として保存する。データは Unicode (UTF-16LE) のテキストファイルとして保存され，各セルはタブ(Tab)によって区切られる。EmEditor で開いて確認しよう [3]。
   ※ Excel のデータをマウスで選択してコピーし，Word 上に「Unicode テキスト」として貼り付けても OK（[編集]→[形式を選択して貼り付け]。単なる「テキスト」では多言語テキストは正しく貼り付けられないので注意)。単純に「貼り付け」をおこなった場合には，フォントなど Excel の書式情報がそのまま Word でも使われるので注意。

2. Word に「エンコードされたテキスト」として読み込む。(Word のオプション設定をあらかじめ変更しておくこと。)

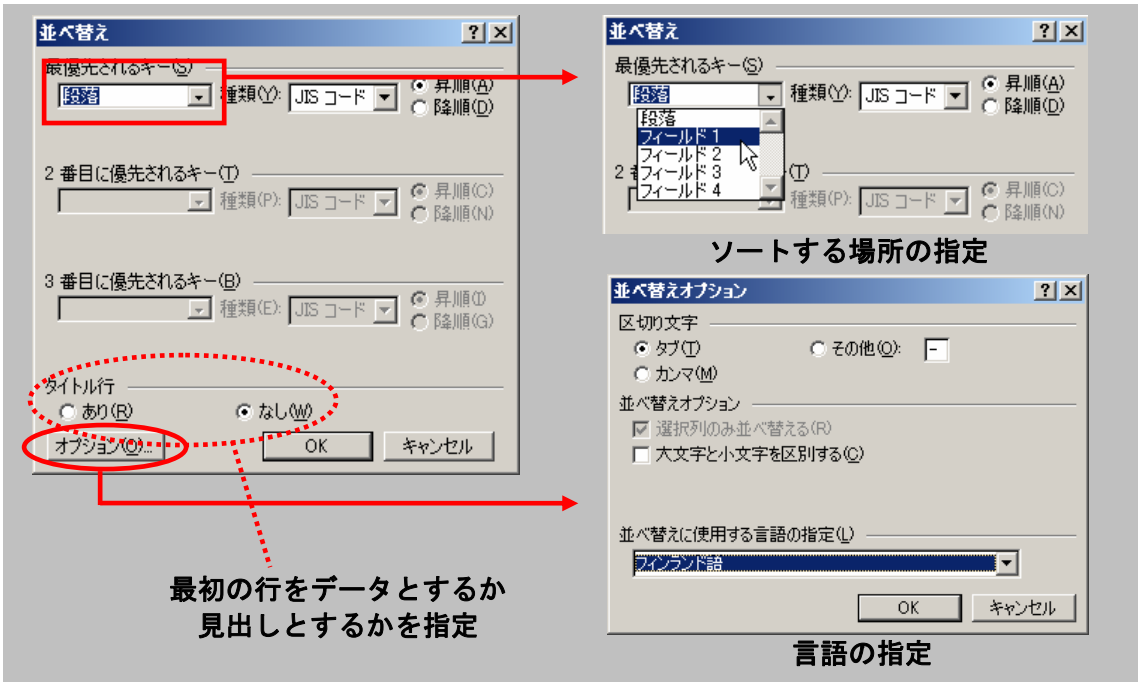
3. ソートする範囲を指定する（文書全体をソートする場合は，指定しなくても OK。見出し行は，ソート時に除外することができる)。

[3] EmEditor の［ツール］→［現在の設定のプロパティ］で画面表示の詳細を設定できる。「記号」タブで「タブ表示」をチェックすると Tab が入っている箇所が青い矢印で表示され，見やすくなる。

4. Word メニューバーから[罫線]→[並べ替え]を選ぶ（右図）。



5. 「並べ替え」ウィンドウの「オプション」ボタンを押し，並べ換えに使用する言語を指定する（下図）。
   - ※ 並べ替えの対象となる単語が行の先頭にない場合，「最優先されるキー」のある列（「フィールド」）を指定する必要がある。
   - ※ 「オプション」では，区切り文字（Excel の「Unicode テキスト」の場合はタブ）の種類，大文字と小文字の区別など，その他の設定も指定できる。



6. ソートが終わったら，結果を確認する。
   - ※ 日本語や中国語のように，読み方が文字と対応していない言語は，文字によるソート自体が難しくなる。そこで，50 音やピンインなどでソートする場合には，読みを別に入力しておき，その読みがなによってソートをおこなうのが確実である（日本語版 Excel の場合，「ふりがな機能」を使った並べ替えもできるが，この場合ふりがなデータは Excel ブック形式のデータにしか保存されない）。

7. 必要に応じ，ソート結果を保存し，さらに編集する（例えば再度 Excel に読み込んで加工したり，Word 文書として保存しデータのレイアウトやフォントを調整することができる）。

   - ※ Word 文書の中でデータを表として整形する場合には，単語データ全体(見出しを含む)を選択し，「表の挿入」ボタンを押すと，  Tab の区切りを自動的に判別して表を作成することができる。

担当：千葉庄寿（schiba@reitaku-u.ac.jp）

※ Word でソートしたデータを再び Excel に読み込み，データを Excel で保存することもできる。Word でソート後，データ部分をマウスで選択しコピーをおこない, Excel のワークシート上に「Unicode テキスト」として貼り付ける（[編集]→[形式を選択して貼り付け]。単なる「テキスト」では多言語テキストは正しく貼り付けられないので注意)。Word のタブ（Tab）ないし表の区切りに従って Excel 上にデータが貼り付けられる。

実習： Kadai サーバの [schiba]→[2006fl]→[No12]フォルダに，フィンランド語，簡体字中国語，繁体字中国語の単語集のサンプルがそれぞれ words_finnish.xls, words_simplifiedchinese.xls, words_traditionalchinese.xls という名前で入っている。これらを使ってソートの練習をしてみよう。

- フィンランド語のデータを Excel のソート機能でソートし, 結果をフィンランド語の正しいアルファベット順 (4 ページ参照) と比べてみよう。
- フィンランド語のデータを「Unicode テキスト」として words_finnish.txt というファイル名で保存し，Word に読み込んで並べ替えてみよう。(Excel の機能を使ってソートした結果を 2 枚目のワークシートに入れておくので，比較してみよう。)
- 簡体字中国語，繁体字中国語のファイルには，2 枚のワークシートにそれぞれピンインありとなしのデータが入っている。Excel のソート機能でソートし，結果を比べてみよう。

最終課題について

日本語以外の言語を 1 つ選択し，Unicode アプリケーションである Excel および Word を使い，各言語の辞書の見出し語順にソートした単語集を作りなさい。とりあげる単語は自分の第 1 外国語や母語を取り上げる場合は最低 100 単語 (150 単語以上を推奨)，第 2 外国語などの場合は 80～100 単語程度とする (特に理由がなければ英語以外の言語を選択すること)。「フランス語の感情と身体動作を表す動詞 100」「ドイツ語インターネット必須用語 100」「中国語による世界の有名人 100 人の表しかた」「Yahoo! China に出てくる電脳語彙 100」「台湾と大陸で異なる表現 100」「『冬ソナ』のキメ台詞(せりふ)で覚える『シブイ』韓国語単語 100」「スペインで体を壊したときの必須単語 100」「英語で学ぶ大相撲用語」など，単語の選定方法や例文の収集方法を工夫しオリジナリティを出すこと。以下の手順 1) ～ 3) に従い，期末試験時に提出すること。

※ 期末試験の詳細については次回説明する。なお，事情により期末試験を受験できない人は，再試験，また課題の提出期日について必ず教員に相談すること。

1) まず Excel を使ってデータを作成する。

- データの構成は以下の条件に従うこと。
  - 単語データは「単語 (語形変化のある言語の場合は辞書形)」,「単語訳(日本語)」,「例文」,「例文訳(日本語)」を含むようにすること。項目は自由に増やして構わないので，各自単語集の内容にあわせて工夫するとよい。
  - 冒頭に必ず見出し (タイトル行) を入れること。
  - 中国語を選んだ場合には，ソートが正しくできるよう，単語とその訳の間に必ず「ピンイン」を加え，第 1 声～第 4 声，軽声をそれぞれ 1-5 の数字で表すこと。
    例： 大学 da4xue2

- 列ごとに適切なフォントを設定すること (標準ではフォントは「MS P ゴシック」という日本語フォントが使われている)。変更したい列のラベルをクリックして列全体を選択し (右図)，フォントを変更する。
- Excel でデータを入力している段階では，フォントの指定以外のレイアウトや文字飾り，罫線などの処理は特におこなう必要はない (以下の 2) で説明するように，Word 上でおこなえばよい)。
- データの入力の際は単語の順番は気にせず，どんどん入力をしていってかまわない。(ソートはデータの入力が終了してから一括して実行すればＯＫ。)
- 作成したデータは words.xls というファイル名をつけて「Microsoft Excel ブック」形式で保存しなさい。

2) Excel で作成した単語データの入力が終わったら，そのファイルを用いてさらに以下の作業をおこない，Word で単語集を完成させなさい。

- Excel を使って作成した単語集データ (words.xls) が完成したら「Unicode テキスト」(= タブ区切りの UTF-16LE テキスト)として保存しなさい。ファイル名は words.txt にしなさい。
- words.txt を Word2003 に読み込み，単語集を**各言語の辞書の見出し語順にソートし**(中国語の場合にはピンインを利用してソート)，データのレイアウトを整えて単語集として完成させ，words.doc として保存しなさい。
  - Word での Unicode テキストの開き方については以下の**補足 2** を参照すること。
  - ページ設定の「印刷の向き」は縦でも横でもよい。その他のページ設定や罫線，フォントの設定は任意だが，ページ数を抑えられるよう列幅などを調整すること。
- (任意) 必要に応じ，日本語でソートしたバージョンも作成し，外国語，日本語の 2 つの言語で閲覧できる単語集を作成してもよい。(日本語でソートする場合には，「単語の訳」以外に「ふりがな」を入れた列を作成しておく必要がある。また項目の順番などを工夫するとよい。) 日本語でソートしたファイルは，ファイル名を適当につけて保存しておくこと。

3) words.doc を印刷したものに以下のように表紙をつけ，あらかじめステープラー(ホチキス)で綴じ，**期末試験時**に提出しなさい。遅延は原則として認めないので注意。

- 課題には表紙をつけ，表紙には「課題タイトル (単語集の内容にあわせ自由につけてよい)」，クラス名，学部学科，学籍番号，氏名，電子メールアドレスを入れること。また，表紙または単語集の冒頭に，作成した単語集の内容に関する解説と，単語の選定方法や例文の収集方法など，利用したデータについて説明すること。
- 印刷は紙資源を節約するため，FinePrint を使い 2 ページを 1 枚に印刷してよい (可読性を確保するため，1 枚あたり 2 ページよりも小さく印刷しないように)。

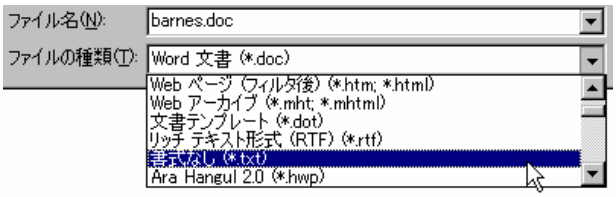
補足 1：自宅で Excel 2000 (2002 以前の Excel) を使って作業をする人へ：

自宅で Excel 2000 (cf. 大学の Excel は 2003) を使って作業をする人は，一部の外国語の表示フォントの指定がうまくいかないことがある。簡体字中国語や韓国語用のフォント (SimSun や Batang, BatangChe など) を指定すると文字化けが起こることがある。そのような場合，文字自体はきちんと入力されていれば，フォントを「Arial Unicode MS」に設定すると表示・編集がうまくできるようになる。(繁体字中国語は MingLiU などを使えば OK。また，簡体字中国語の場合は，ユニコード対応フォントの別名 N SimSun を指定すると，文字化けが回避できるようである。)

担当：千葉庄寿（schiba@reitaku-u.ac.jp）

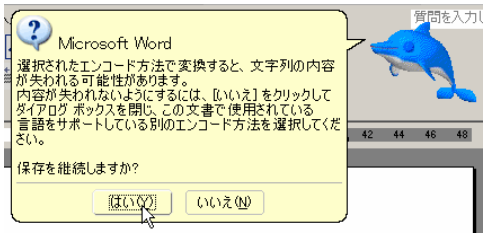
補足 2：Unicode を用いたテキストファイルの作成・保存と編集—Word の場合

■ Word 文書をテキストファイルとして保存



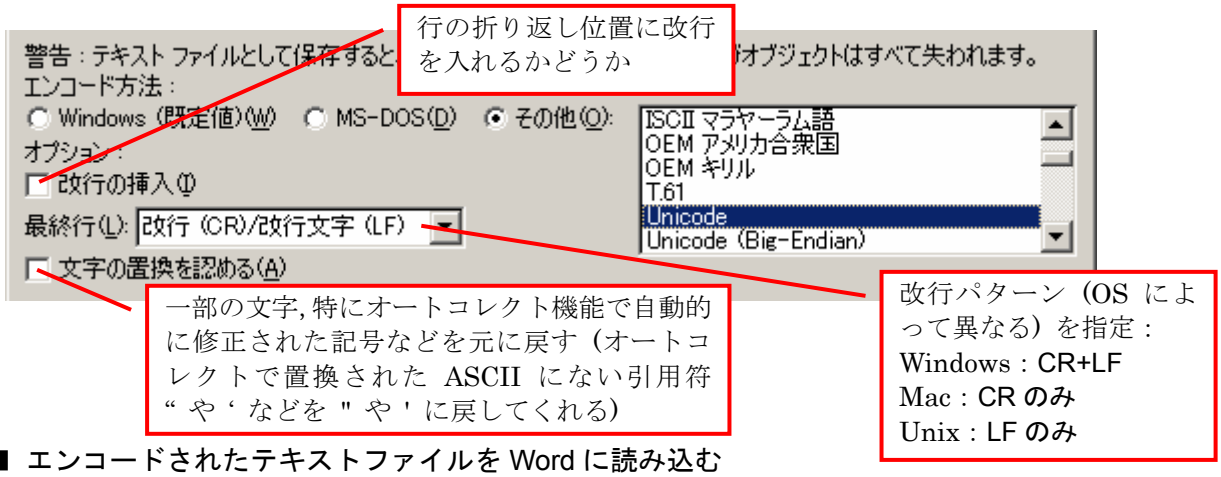
Word 2003 には，[ファイル]→[名前を付けて保存]を選択し，ファイルの種類として「書式なし(*.txt)」を選ぶことで，ファイルをテキストファイルの形式で保存する機能がある。

「書式なし」を選んで保存したテキストは，文字エンコード方式を指定して保存することができる。「エンコード方法」として「Windows (規定値)」(大学 PC の場合は日本語 Shift JIS)，ではなく「その他」を選び，具体的な文字エンコード方式を選択する。

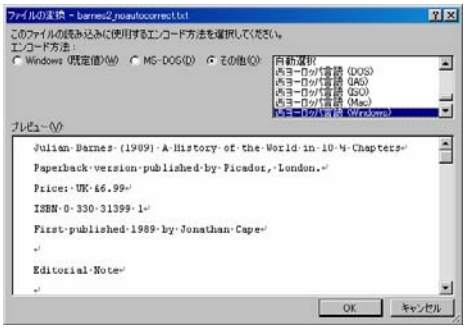


なお，指定した文字エンコード方式が文字を正しく保存できない場合，右図のような確認メッセージが表示される。

「ファイルの変換」画面のオプションの詳細は以下の通り：



■ エンコードされたテキストファイルを Word に読み込む



Word2003 には，テキストファイルの読み込み機能がある。これを使って，さまざまなエンコード方式で作成されたテキストファイルを Word に読み込み編集することができる。

1. [ファイル]→[開く]を選択する
2. 読み込むファイルを選択する。[ファイルの種類]を「テキストファイル(*.txt)」ないし「全てのファイル(*.*)」に指定し，テキストファイルを選択する。
3. 「ファイルの変換」ダイアログでエンコード方法を選択する。「エンコード方法」を適切に指定し，正しい文字エンコード方式を選ぶ（変換内容はプレビューで確かめることができる）。
4. 正しく表示できたことを確認したら，「OK」ボタンをクリックして文書を開く。

注意：Word でテキストファイルを開くと，文字エンコード方式によってはフォントに Arial Unicode MS などが指定され，行の高さが通常より高くなることがある。ファイルを Word で編集して印刷する場合は，フォントを適切なものに変更してからおこなうとよい（文書を再度テキストファイルとして保存する場合，フォント情報は無視されるので注意）。