語学教育素材作成を支援する多言語分析ツールの開発

成定久美子† 佐野 洋‡
東京外国語大学 †大学院地域文化研究科 ‡外国語学部
〒183-8534 東京都府中市朝日町 3-11-1
e-mail: narisada.kumiko.rra@tufs.ac.jp

概要

本稿は，語学教育用の教材作成支援ツール(KOTOEMON)について述べる．これは日本語教育のための素材作成支援ツール(CLTOOL)を改良したもので，海外で日本語教育に従事している外国人が日本語教育用素材を効率的に作成するためのソフトウェア支援ツールである．さらに発展させ，外国語教育に従事している教師が外国語教育で使用する教育素材を効率的に作成するためにも利用できる．このツールを利用して作成した教育素材を利用することで，いわゆる母語話者の言語直感に頼るのではなく，言語の運用事実に基づいた用例を教育教材に反映することができる．

本ソフトウェアは，多言語の処理に対応するためUnicodeに対応した文字処理機能を有する．また，海外における日本語教育での利用を考慮して，ユーザインタフェースに表示されるメニュー等のテキスト表示を多言語化した．さらに，語学教育に携わる人が文系出身者であること勘案し，文型検索のための正規表現の入力をはじめとする処理手続きの視覚化を図り利用しやすいソフトウェアとした．

1. はじめに

本稿は，語学教育用の教材作成支援ツールについて述べる．これは日本語教育のための素材作成支援ツール(CLTOOL)[2][6]を改良したもので，海外で日本語教育に従事している外国人が日本語教育用素材を効率的に作成するためのソフトウェア支援ツールである．さらに発展させ，外国語教育に従事している教師が外国語教育で使用する教育素材を効率的に作成するためにも利用できる．

教育素材の作成は，非常にコストのかかるプロセスである．数多くの資料にあたり，その中から学習段階に適切であると考えられる素材を抽出しなければ高い学習効果は得られない．しかし教育者は，その言語を専門に教育しているとはいえ，その読書範囲や生活空間によって学習用例の抽出には偏りがあることも否めない．KOTOEMONを利用することで，広範な表現の分布を高速に検索し，適切な表現例を効率的に見つけ出すことができる．さらに，いわゆる母語話者の言語直感に頼るのではなく，言語の運用事実に基づいた用例を教育素材に反映することができるだろう．

KOTOEMONは東京外国語大学のTUFS言語モジュール[1]の作成を支援するために，対象となる17言語(英語，ドイツ語，フランス語，スペイン語，ポルトガル語，ロシア語，中国語，朝鮮語，モンゴル語，インドネシア語，タガログ語，ラオス語，カンボジア語，ベトナム語，アラビア語，トルコ語，日本語)に対応させることを最終目標としている．多言語処理を可能にするため，Unicodeに対応した文字処理機能を有する．

想定する利用者は，外国語教育に携わる人（人文系出身者）である．コンピュータへの恐怖心の払拭と利用の便とを考慮し，コマンド手続き的なインターフェースを改めて，処理手続きの視覚化を行った．例えば，文型検索のための正規表現の入力では，直接正規表現を入力させないで，予め用意した正規表現を選択するだけで済むようにした．また，海外における利用を考慮して，ユーザインタフェースに表示されるメニュー等のガイドのためのテキストを多言語化した．

Development of a Multilingual Analysis Tool for Preparation of Language Teaching Materials
K. Narisada, H. Sano
Tokyo University of Foreign Studies

[1] TUFS言語モジュールとは，文部科学省が推進する21世紀COE（Center of Excellence）プログラムに採択された「言語運用を基盤とする言語情報学拠点」の研究成果を活かして開発した，新しいインターネット上の言語教材である．

2. 語学学習素材作成支援ツール

CLTOOL は，既存の日本語研究用ソフトウェア[8]を，適応保守の工程を経て教育教材開発支援用に改変・開発したものである．Perl/Tk[2]を利用したユーザインタフェースを実装し，文字・語彙調査だけではなく，複文・連文調査にも活用できるように機能を拡張した[4][5]．このツールを利用して作成した教育素材を利用することで，いわゆる母語話者の言語直感に頼るのではなく，言語の運用事実に基づいた用例を教育教材に反映することができる．

しかし CLTOOL では，(1)日本語 OS の利用環境を前提としていたため，海外のコンピュータ環境ではメニュー表示が文字化けしてしまう，(2)同じ理由で，処理対象データが日本語(シフト JIS)コードのテキストに制限される，といった問題点があった．これらを解消するために，本研究では CLTOOL を改良ローカライズして語学教育用の教材作成支援ツール(KOTOEMON)を作成した．

KOTOEMON では，Unicode を使用することでメニュー等の文字化けを解消した．また，半自動的に文字コードを検出する機能を追加し，どのような文字コードのテキストにも対応できるようにした．

3. KOTOEMON

KOTOEMON は，日本語分析に特化していた CLTOOL をローカライズしたものである．ローカライゼーション(多言語化)とは，アプリケーションをある特定の言語または地域の稼動条件に適応させることである[10]．本稿では，表示されるメッセージを利用者の母語に翻訳したり，各言語を処理するための各種辞書や定義ファイルを用意したりすることを「言語的側面でのローカライゼーション」という．また，利用者に合わせて情報や機能を変更することを「文化的側面でのローカライゼーション」という．

3.1 言語的側面でのローカライゼーション

3.1.1 分析対象言語の拡張

日本語以外の言語も扱いたいという要望を受け，KOTOEMON では分析対象言語の拡張を行った．内部処理コードに Unicode を使用しているので，例えば KWIC[3]機能ではアラビア語，中国語をはじめとする 66 ヶ国語を検索することが可能である．中国語（簡体字）の検索結果の例を図 1 に示す．

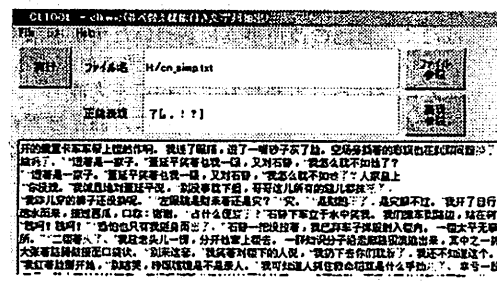

図 1. 中国語の検索結果の例

分析対象言語の拡張に伴い，ツールに含まれる情報や機能を，(1)特定の言語と結びつかないように抽象化された機能と，(2)対象言語固有の機能とに分類し，モジュール化した．これにより対象言語や表示用言語を追加したり，特定の言語特徴に特化した機能を追加したりすることが容易にできるようになった．また，利用者は必要なモジュールのみを選択してインストールできるため，システムの巨大化を防ぐことができる．

3.1.2 データ処理

KOTOEMON には各言語を分析するための各種辞書と，正規表現や文区切り記号を定義したファイルが用意されている．利用者は対象の分析言語モジュールをインストールすることでその言語で記述されたファイルを分析することができる．データ処理の概念図を

[2] Perl 言語を用いて Window 上で動作する GUI(グラフィカル・ユーザ・インターフェース)アプリケーションソフトを構築するためのライブラリ(ソフトウェア部品)．

[3] KWIC とは「Key Word In Context」の略で，検索文字列を中心にして前後の文脈とともに表示したものをいう．

図 2 に示す.

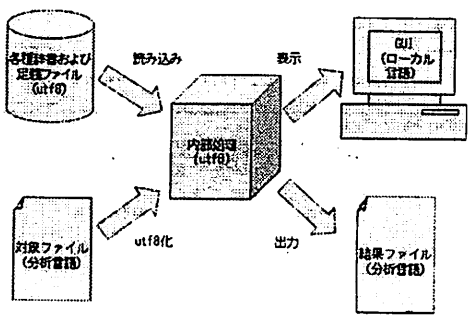

図 2. データ処理の概念図

分析対象言語に関わる各種言語および定義ファイル(文区切り記号や正規表現)と，メニュー表示等に関わるデータのファイルはそれぞれ図 3，4 のように構成されている.

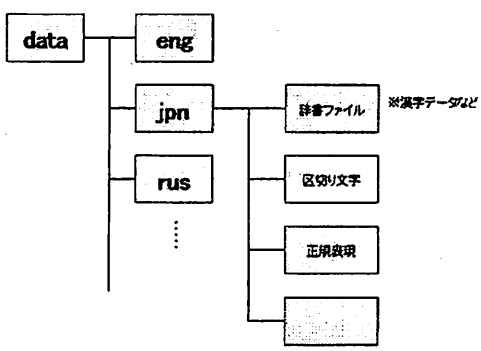

図 3. 分析言語に関するデータの構造

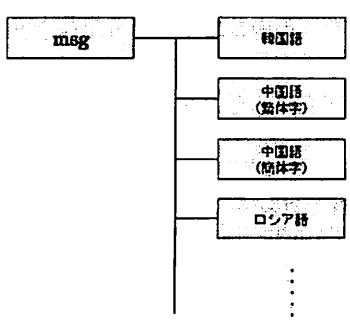

図 4. メニュー等表示に関するデータの構造

3.2 文化的側面でのローカライゼーション

3.2.1 メニュー表示

ローカライゼーションはアクセシビリティを向上する上で非常に重要な条件の一つとなってきている．ある調査によると，母国語で書かれたサイトから商品を買うという小売客は，そうでない客の約三倍にも上るという[14]．ユーザインタフェースで表示されるテキストは，利用者との意思疎通のためのコミュニケーションの道具である[1]．そこで本ツールでは，ユーザインタフェースの全ての表示を多言語化することでソフトウェアへの心的障壁を低くすることに注力した.

利用者はメニューに表示される言語と分析する言語を，PC を再起動させることなく簡単に切り替えることができる(図 5).

図 5. 表示言語と分析言語の選択

もっかのところ，日本語，中国語(簡体字・繁体字)，ロシア語，韓国語，アラビア語，英語の 7 つの表示モードが選択可能である．中国語(簡体字)の表示例を図 6 に示す.

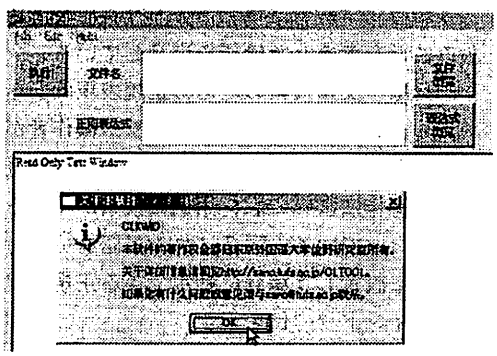

図 6. 中国語(簡体字)の表示例

ローカライズに際して，メニュー表示されているテキストを，元の日本語表示を直裁に翻訳すればよいわけではない．言語によって

は，ある用語の訳語が存在しない場合もある．また，翻訳することで表示に要するスペースが変わってしまう場合もある．例えば，英語のテキストをフランス語に翻訳すると 15%～20%，ヒンディー語に翻訳すると 80%，テキストの分量が増加する[12]という[14]．そのため，ユーザインタフェースのデザイン設計をおこなう際には，こうした翻訳語のためのスペースを確保しておく必要がある．

また，翻訳対象となるテキストは比較的短い文章や語句が多く，外部化した際に文脈(文字列の出所)が不明になり，訳を特定できない場合がある[1]．例えば，「ファイルが見つかりません」というエラーメッセージをアラビア語に翻訳する場合，「誰が見つけられないのか」という人称の特定が必要になるのである．

本稿では，日本語教育素材作成支援ツールのマニュアル[7]を使用し，機能と表示画面をあらかじめ翻訳者に提示するとともに，分かりにくい語句を説明文に直したファイル[4]を用意することで，翻訳する語句の「文脈」を伝えるようにした．

3.2.2 文字パターン入力支援

本ソフトウェアの特徴は，外国語教育に従事している教師の教育素材作成を効率的に支援するために，各言語の文法的な特徴に合わせた機能の変更・追加を行ったことにある．とりわけ語学教育に携わる人が文系出身者であることを考慮した機能を用意している．

例えば，文型検索のための正規表現はあらかじめ言語ごとに用意されており，操作者が必ずしも正規表現を直接入力しなくても，直感的な操作によって教材用例の検索ができるようになっている．図 7 にロシア語を分析対象言語に設定した場合に利用できる正規表現参照ウィンドウの例を示す．

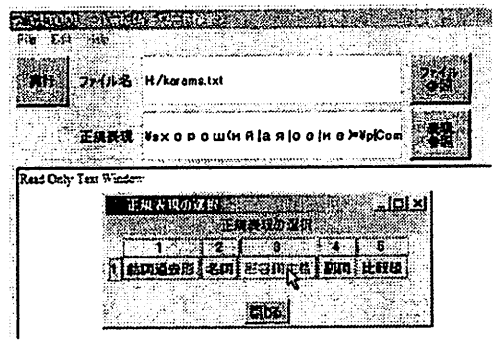

図 7. ロシア語の正規表現参照ウィンドウ

ロシア語の正規表現選択ウィンドウでは，[形容詞主格]を押すと，入力ボックスに "¥sхорош(ий|ая|ое|ие)*¥p{Common}" が自動的に入力される[5]．同様に[動詞過去形]を押すと"¥sчитал(а|и|о)*¥p{Common}"が入力されるのである．

3.2.3 文字コード変換支援

KOTOEMON は内部処理コードに Unicode を用いている．さまざまの文字コードを統一的に扱うには都合がよい．一方，本ソフトウェアを利用する側から考えると，処理対象の言語の文字コードを正確に把握して，Unicode に変換する必要性が生じることになる．とりわけ文系出身者にとって，このような技術的な理解を求める一連の操作は負担感を生じさせる要因になっている．この技術負担を軽減し，直感的な操作によって文字コードの変換を容易に行うために，KOTOEMON では，文字コード変換に際して，処理対象のテキストを，その言語を表現可能な文字コードに変換して表示する機能を有している．文字コード変換の例を図 8 に示す．

---

[4] 例:「KWIC 文脈内分割」→「文脈からキーワードを切り離す」のように，名詞句ではなく文に直すことで，翻訳箇所の文脈を理解してもらえるようにした．

[5] ¥s は空白文字(スペース，タブ，改ページ，復帰，改行) 1 文字を表す．

¥p は Property 名を表す．Unicode では，膨大な数からなる文字を扱いやすいように，文字集合を Character Databas，Script，Script Block の 3 つの視点から定義し分類している．Perl では，これらの Property を持つ文字を ¥p{} という書式でまとめて表現することが出来る．例の¥p{Common}はスペースや句読点などの文字集合を表している．

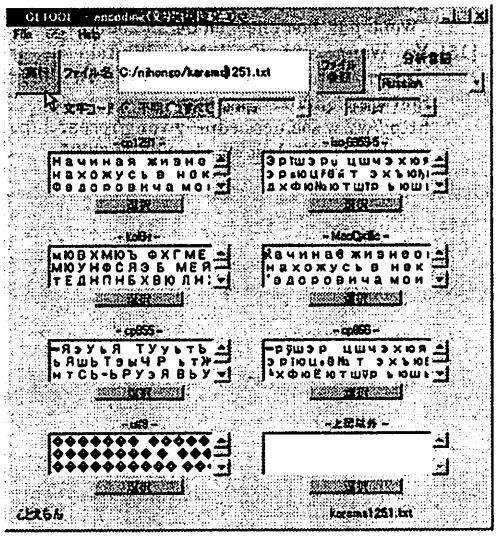

図 8. 文字コード変換の例

利用者は，表示されたものの中から正しく文字が読めるものを選択することにより，文字コード名を正確に知る必要なく，例えば，koi8[6]や cp1251[7]などから Unicode にテキストの文字コードを変換することができる．

3.2.4 特定言語に特化した機能の追加

その他にも，日本語や韓国語など，ローマ字が限定的に使用されるなど書記体系が混在する言語については，その特徴に合わせた機能を追加した．[4][5]では文字の種類ごとに統計処理を実施する機能が開発された．文字種は日本語に特徴的な性質で，テキストの文字種の分布を知ることは，そのテキストの複雑さや読みの容易さなどに連関しているのである．日本語教育にも応用可能な機能である．KOTOEMON では，さらに日本語教育の教材開発の点を重視し，指定された漢字に読み仮名を付与する機能を追加した．

言語が異なると使用される句読点記号も異なる．英語の疑問文では文末に付与される疑問符は"?"だが，ドイツ語では，英語のセミコロンに似た記号が使われる[15]．そのため KOTOMEMON では，言語ごとの句読点記号を選択して指定できるようにして，言語の違いによる区切りの多様性にも対応している．

こうした工夫にもかかわらず，さまざまな言語に対して詳しい検索を実現するには不十分である．例えば，アラビア語のような right-to-left 入力の言語を詳細に検索するためには，2 方向入力に対応したインターフェースおよび正規表現の検討が必要である．また，検索結果をソートする場合，その順番は言語や国によって異なる．日本語や中国語ではコード順だけではなく，発音，部首，画数で並び替える方法を検討する必要がある[15]．もっか，こうした言語依存の細かな処理機能の開発を進めている．

4. おわりに

現在，東京外国語大学の TUFS 言語モジュールに対応できるよう，ユーザインタフェースに現れるテキスト表現の翻訳および分析対象言語の追加を進めている．今後は言語ごとの文型特徴を調べて，正規表現を充実させていく予定である．

また，現段階で完成している機能を語学教育者や研究者に実際に利用してもらい，言語に固有な性質を分析することで一部言語対応のモジュールを実装することも検討している．さらに改善を重ねたい．

5. 謝辞

本研究は平成 14-16 年度文部科学省科学研究費(基盤研究(B)(2))「全電子化検定済み教科書データの解析と大規模日本語コーパスの構築」(研究代表者　佐野洋)および平成 17 年度文部科学省科学研究費(基盤研究(C))「教材制御の枠組みに基づく英語 e-Learning の研究開発」(研究代表者 馬場 彰)の助成を受けた．

また本論文をまとめるにあたり，有用な御指導とコメントを頂きました査読者の方々に感謝いたします．

[6] koi8: ロシアにおける事実上の標準文字コード．各種記号やキリル文字圏の特殊文字をサポートしている．

[7] cp1251: Microsoft Windows におけるキリル文字標準セット．CP とはコードページの略である．

参考文献

[1] 加藤直孝・有澤誠, Conceptualization of Program Integrated Information, 自然言語処理研究会報告, (社)情報処理学会, 2005.

[2] 佐野洋,「ソフトウェア再利用による日本語研究のための分析ツールの開発」, 電子情報通信学会総合大会公演論文集, IEICE 総合大会, 2003.

[3] 佐野洋,「日本語教育素材開発のための支援ツール」, PC カンファレンス論文集, pp413-416, CIEC(コンピュータ利用教育協議会), 2003.

[4] 佐野洋,「日本語学習素材作成のための日本語処理ソフトウェア」, CIEC(コンピュータ&エデュケーション)会誌 2003 年 VOL.15(8 頁), 2003.

[5] 佐野 洋,「日本語教育素材作成のための日本語分析ツールの開発」, 情報処理学会, コンピュータと教育
研究会報告(CE-70), 講演論文集, pp.21-26, 2003.

[6] 佐野洋, 幸松英恵「ソフトウェア再利用による語彙調査用ツールの開発」, 言語処理学会第 9 回年次大会公
演論文集, 言語処理学会第 9 回年次大会, 2003 .

[7] 佐野洋,『Windows PC による日本語研究』, 共立出版, 2003.

[8] 中野洋,「パソコンによる日本語研究法入門」, 笠間書院, 1996.

[9] Arle Lommel 編,『ローカリゼーション業界への手引き 第 2 版』, SMP Marketing & LISA, 2004.
http://www.lisa.org/products/primers/primer2_jp.pdf, last checked 07/14/2005.

[10] Sybase® Adaptive Server® Enterprise, "バージョン 12.0 Sun Solaris 2.x (SPARC) 版 インストール・ガイド",
http://download.sybase.com/pdfdocs/asp1200j/jaseigso.pdf, last checked 07/14/2005.

[11] Gregory E. Kersten, Mik A. Kersten and Wojciech M. Rakowski, "Software and Culture: Beyond the Internationalization of the Interface", Journal of Global Information Management, 10(4), 86-101, Oct.-Dec. 2002.
http://kerstens.org/mik/publications/softwareAndCulture-jgim2002.pdf, last checked 07/14/2005.

[12] Jack D. Grimes, Deborah J. Knoles and Mark E. Davis, "Creating global software: text handling and localization in Taligent's CommonPoint application system - Taligent Inc", IBM Systems Journal, June, 1996.
http://www.research.ibm.com/journal/sj/352/davis.pdf, last checked 07/14/2005.

[13] Nick Symmonds, "Internationalization and Localization Using Microsoft .NET", 2002. XIX, Edition.
http://www.programmersheaven.com/other/BookSamples/pdf/Symmonds.pdf,
last checked 07/14/2005.

[14] Rosann Webb Colliins, "Software Localization: Issues and Methods", *in the 9th European Conference on Information Systems - ECIS 2001*, (Bled, Slovenia),
http://is.lse.ac.uk/Support/ECIS2001/pdf/003_Webb.pdf, last checked 04/28/2005.

[15] Yu Wang and E. James Whitehead, Jr., "International Accessibility of Open Source Software", 2001.
http://www.soe.ucsc.edu/~ywang/research/papers/oss01.pdf, last checked 04/28/2005.