フィールド言語学者のニーズに合った
多言語処理ツールの試み

小野　智香子 *・鈴木　麗璽 **・松村　一登 ***

日本学術振興会 特別研究員／東京大学大学院人文社会系研究科 *
名古屋大学大学院人間情報学研究科 博士課程 **
東京大学大学院人文社会系研究科 教授 ***
ono_suz_mac@kmatsum.info

背景

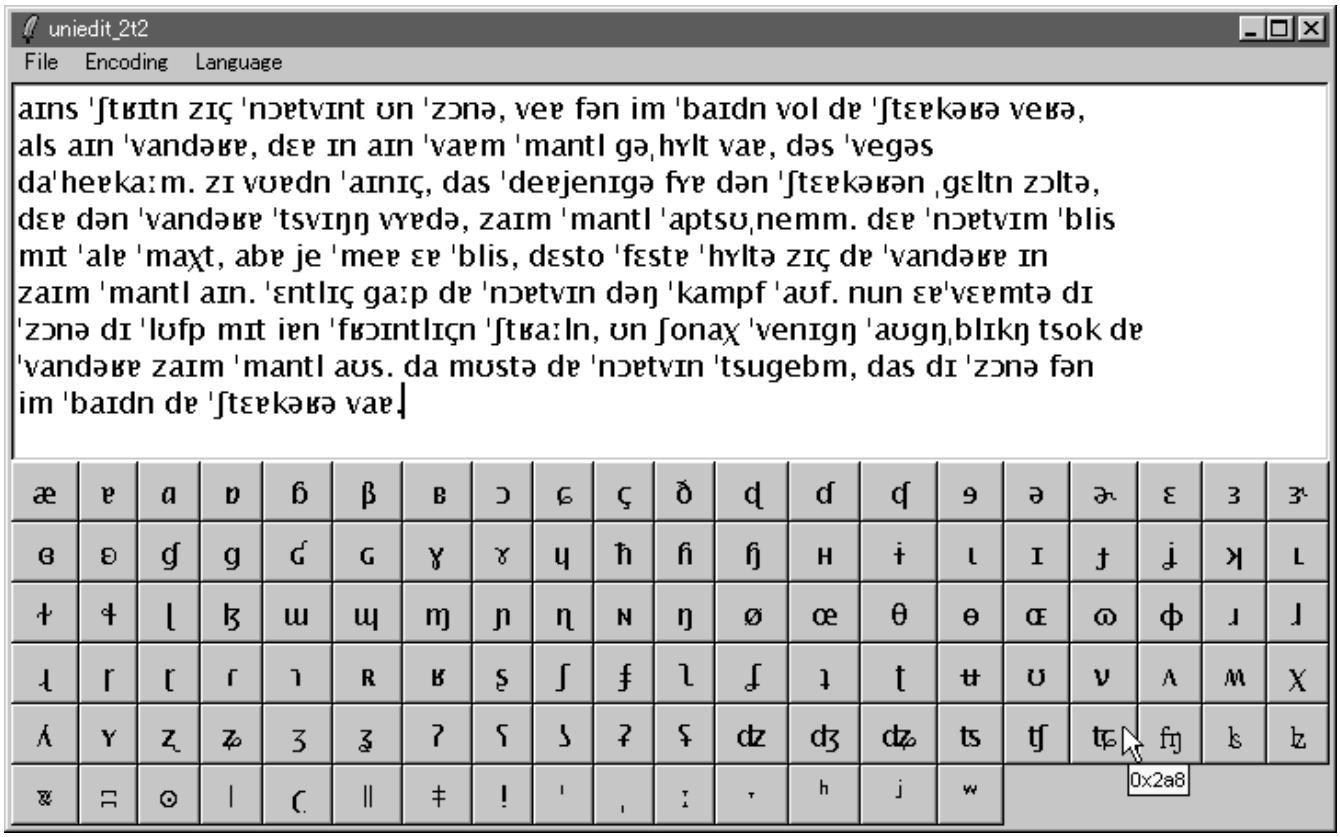
いわゆる文字のない言語や特殊な文字体系を用いる言語のデータをコンピュータで処理する場合，ASCII 文字やその組み合わせで代用したり，独自に外字やフォントを作成したりして対応してきた研究者は多い。しかし，独自に作成された外字やフォントはそれらを組み込んだコンピュータでしか利用できず，また ASCII 文字で代用した場合も文字の組み合わせと音の対応が研究者によってまちまちであるため，研究者の間でデータや研究成果のやりとりを行う際に，文字化けが起こったり，音声記号の代用として使われている文字と音声記号との対応関係が推定できない場合があるなど，問題が多い。

紙のメディアを媒介としない電子的な文書交換が急速に一般化しつつある現在，音声記号や特殊な文字を含む言語データの確実なやりとりを紙のメディアに依存して行うことの限界が誰の目にも明らかになりつつある。いわゆる文字のない言語のデータの記録の手段である音声記号のエンコーディングが統一されていない状態のまま，コンピュータを用いて文書が作られ続けるなら，早晩，言語学者たちののハードディスク上に蓄積されている言語データが，そっくり文字変換処理の対象となって，膨大な時間と費用を費やさなければならない事態になるのは必至であり，最悪の場合，紙のメディアに出力してないハードディスク上の言語データの一部が，著しい文字化けによってもとの姿に復元できなくなってしまう可能性さえある。

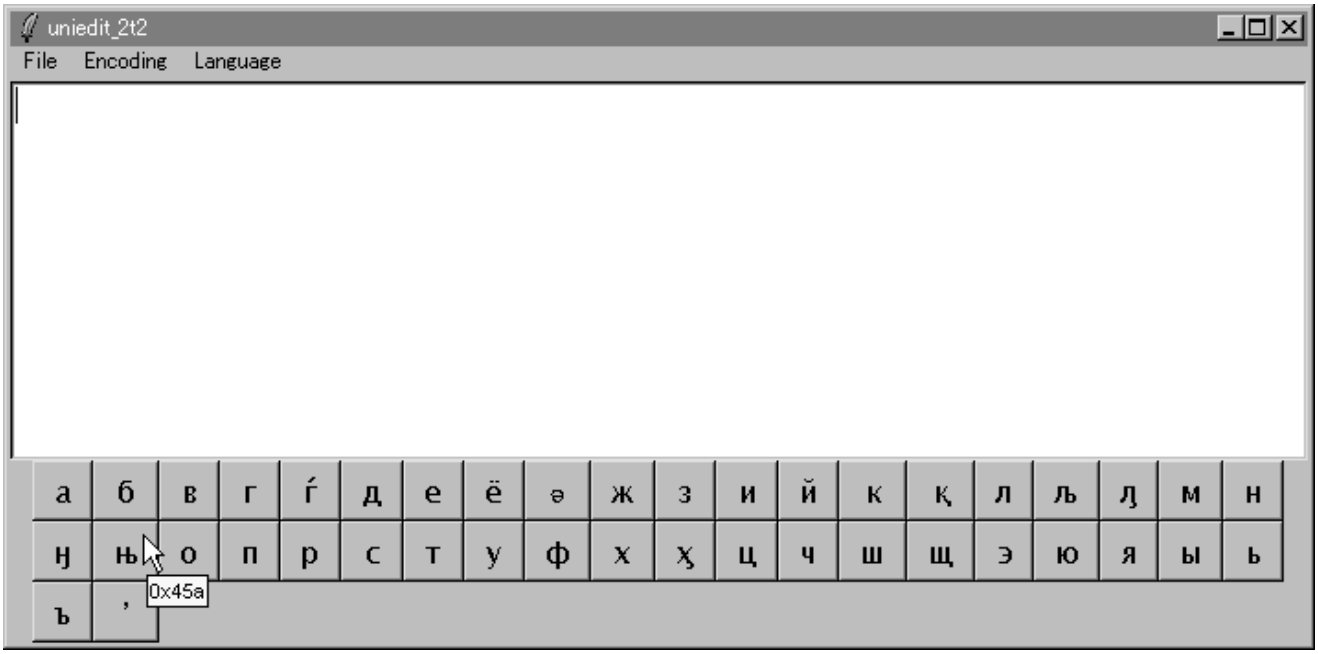
この問題の解決方法は，ただ一つ，特定のソフトウェアや特定の文字フォントに依存しない多言語処理方式を皆が共有することである。コンピュータによる多言語処理のために，**Unicode**（ユニコード）と呼ばれる国際規格が開発されている。世界中の様々な文字体系を一つの巨大な文字セットとしてまとめ，1つ1つの文字に一義的な文字コードを与えることによって，文字と文字コードの間の曖昧性をなくすことを目指す Unicode のような規格が普及すれば，文字化けというリスクがなくなり，コンピュータによる多言語処理の効率が格段に高まることが期待される。

すでに，コンピュータの OS のレベルで の Unicode による文字処理が一般化しているほか，文字テキスト処理のためのエディターやワードプロセッサも，Unicode による文書保存・読み込みが可能な仕様になってきている。言語学者のニーズに即していえば，**IPA** 独自の音声記号は，Unicode 規格の IPA extensions という領域に登録されていて，たとえば Windows なら，デフォルトでインストールされる Lucida Sans Unicode という名前のフォントに IPA extensions の領域の文字や記号がほぼ含まれている。また，日本語 Windows の標準の漢字フォントにもたいていの IPA 記号が含まれている。

IPA などの特殊な文字がデフォルトでカバーされるようになってきていること自体は朗報なのだが，この種の文字を実際に文書の中で使おうとすると，簡単に入力できる手段が事実上提供されておらず，せっかくの Unicode が活用できないという壁にぶつかる。Windows 2000 を例に取れば，「アクセサリ」フォルダー内の「システムツール」に「文字コード表」と呼ばれるユーティリティーがあり，日本語 FEP の「記号



【図1】IPA 音声記号のソフトキーボード
（テキストは【文献3】p.88 のドイツ語の「北風と太陽」の音声表記）



【図2】チュクチ・カムチャツカ諸語入力用のキリル文字ソフトキーボード

入力」「コード入力」などと呼ばれる機能に対応する機能を Unicode 文字用にも提供する。「文字コード表」は，フォントを選ぶとそのフォントにグリフが登録されている文字の一覧を表示し，文字をマウスで指定すると「U+0260: Latin Small Letter G With Hook」のように文字コードと名称を教えてくれるほか，文書に貼りつけることができる。しかし，日本語 FEP の補助として行う「記号入力」「コード入力」に相当する機能しかない「文字コード表」を使って，IPA の音声記号で転写された言語データをまとめて入力するのは至難の技である。

このような現状をふまえて，Windows や Mac に最初から付属している Unicode のリソースを，一般のユーザ，とりわけフィールド言語学者にも利用できるものにするための最小限の条件を考えてみると，次の2点に集約すると考えられる。

I. いちいち Unicode の文字コードを打ち込まなくても，音声記号や特殊な文字を入力できるツール。具体的には，コンピュータのキーボードとも連動し，マウスで文字を選んで入力できるソフトキーボードで，ユーザのニーズに合わせて，Unicode 文字集合の適当な部分集合を「サーミ語ラテン文字表記」のような名前で登録でき，必要に応じて文字セットが切り替えられる仕様になっているのが理想的である。できれば，このツールは特定の OS に依存しないものであることが望ましい。

II. Unicode に登録されている任意の文字・記号を画面上に表示し，プリンタから出力するための Unicode 規格に対応した文字フォント。このフォントは，誰でも無料で利用できるものでなければならない。もちろん，Windows や Mac に出荷時から付属している Unicode 対応のフォントに収録されている文字・記号だけで十分な場合は，このフォントを使わずに済ますことができる。

この2つの条件が満たされるなら，いわゆる文字のない言語の資料を音声記号で記録している言語学者たちの苦労が半減するだけでなく，標準ではサポートされない特別な文字体系をもつ言語の文書作成にともなう技術的な障碍を大幅に軽減し，言語学者の研究成果の現地還元の促進や，少数言語を母語とするコミュニティーにおける出版事情の改善に多大な貢献ができるはずである。

Unicode 対応の汎用ソフトキーボード

フィールドで採集したデータをそのまま Unicode を使って入力できるツールがあれば便利である。この要請に応えるために，汎用的な Unicode 文字入力用のソフトキーボードの開発を試みた。このソフトキーボードは，スクリプト言語 TCL/Tk に関する情報ソース・サイト Wiki において Richard Suchenwirth 氏が公開している「**A Little Unicode Editor**」のソースコードに，適宜改良を加えたものである。

このソフトキーボードは，入力したい言語（文字セット）を選択すると，ウィンドウ下のキーボードがその配列に切り替わり，マウスでキーを押すと，その文字がウィンドウ上のテキスト領域に入力される。なお，キーの上にマウスポインタを移動させると，その文字の Unicode でのコードがポップアップして表示される。テキスト領域上でマウスの右ボタンをクリックすると，ポップアップメニューが現れ，テキストの選択範囲をコピーまたは全体をコピーすることができ，その後他のアプリケーションにペーストし，利用することができる。**【図1】**・**【図2】**参照。

このソフトキーボードの特徴は，キーの配列を独自に設定可能なことである。スクリプトと同じディレクトリに存在する設定ファイルには，キー配列の名称とその配列を表す Unicode の文字コードが羅列されている．これをユーザが任意に編集することで，用途に応じたキーボードを作成することができる。

ユニコード対応フォント

Unicode を用いてテキストファイルを作成しても，それに含まれる文字のグリフを含んだ Unicode 対応のフォントを皆が利用できなければ，どの環境においても正しく表示することはできない。しかし，現在，改変等も含め完全に自由に利用可能なUnicode 対応フォントはごくわずかである。また，既存のフォントの多

くは，言語毎に限られたグリフしか登録されておらず，この場合複数のフォントを明示的・非明示的に切り替えて利用することになり，不便である。そこで，公開されている改変再配布自由なフォントを収集し，統合することで，多くの Unicode の文字を表示可能なフォントを開発する **Free UCS Outline Fonts** プロジェクトが進行中である。このプロジェクトでは FreeSans, FreeSerif, FreeMono という3種類のフリーなフォントを開発中であり，欧米・ロシア・アジア圏等の広範囲な文字を収録予定である。しかし，当初の予定では漢字が含まれておらず，また，現状では IPA の音声記号の整備が不十分である。そこで，われわれは改変再配布自由な日本語フォントと，この FreeSerif フォントを統合し，辞書ツールで用いている全ての言語の文字を表示可能なフォントを開発中である。

多言語対応の語彙検索ツール

試みとして，発表者のひとりが参加して作成したチュクチ・カムチャツカ語族（チュクチ語，アリュートル語，コリャーク語，イテリメン語）の比較基礎語彙集（1000語）のデータ【文献1】(以下「比較語彙集」) を用いて，検索語の入力と検索結果の出力が，原語，日本語，英語，ロシア語のいずれでも可能な，多言語対応の語彙検索ツールを作成してみた。

「比較語彙集」に収録された語彙 1000 語は，【文献2】に準拠して選定されたもので，【図3】左のような英語・日本語と IPA 表記のデータだけが並んだページが前半部，【図3】右のようにロシア語とキリル文字表記のデータだけが並んだページが後半部としてまとめられている。巻末にチュクチ・カムチャツカ語族の語形のアルファベット順索引がついているが，本体は見出し語を意味分類にもとづいて並べているため，辞書風の検索には適さない。

0226	person, man, one ひと(人)	0226	человек
Ch	ʔorawetɬʔan	Ч	ъораъВэтԓьан
A	ʕujamtawilʔən	А	ГуйамтаВилъын
K	ʕujemtewilʕən	К	ГуйэмтэВилГын
NI	cˀamzanɬχ	И(с)	чˀамзандχ
SI	cˀamzanɬχ	И(ю)	чˀамзандχ

【図3】「比較語彙集」のデータ形式

0226 ひと(人)
person, man, one
человек
ʔorawetɬʔan / оˀравэтльан
ʕujamtawilʔən / гˀуямтавˀилъын
ʕujemtewilʕən / гˀуемтэвˀилгˀын
cˀamzanɬχ / чˀамзандχ
cˀamzanɬχ / чˀамзандχ

【図4】「辞書検索ツール」のデータ形式

「辞書検索ツール」を作成するにあたっては，「比較語彙集」でキリル文字表記を一種の音声表記として用いていたのを，正書法の表記に置き換えたうえで，IPA 表記の語形とキリル文字表記の語形を【図4】のように一箇所にまとめて，一回の検索で両方の表記が得られるようにした。「辞書検索ツール」の語彙デー

タは，UTF-8 でエンコードされ，各項目がタブで区切られたファイルになっている。

「辞書検索ツール」は，チュクチ語を IPA表記で入力し，検索結果を日本語，英語，ロシア語等々で表示したり（**【図5】**），ロシア語を入力して，検索結果のチュクチ・カムチャツカ諸語を IPA 表記とキリル文字の正書法の 両方で同時に表示する（**【図6】**）など，ひととおりの検索方法が可能である。また，単語全体と一致する場合に限定するか，単語の一部と一致する場合も含めるかどうかも指定することができる。

開発言語について

本研究では，ウィンドウやボタンなど，グラフィカルなインターフェイス (GUI) を基本としたアプリケーションを比較的容易に作成することができるスクリプト言語の一つである TCL/Tk をソフトウエア開発のための言語として採用した。TCL/Tk は，オープンソースによって開発されているフリーな言語であり，Windows や Mac の環境で，Unicode を正しく処理・表示できる。

参考文献，参考 URL

語彙データ

1. 呉人恵・編『チュクチ・カムチャツカ語族比較基礎語彙集：1』科学研究費補助金特定領域研究（A）「環太平洋の『消滅に瀕した言語』にかんする緊急調査研究」，2001.
2. 『アジア・アフリカ言語調査票・上』東京外国語大学アジア・アフリカ言語文化研究所，1966.

国際音声字母 (IPA)

3. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge University Press, 1999.
4. The International Phonetic Association, **http://www.arts.gla.ac.uk/IPA/ipa.html**.

Unicode 規格

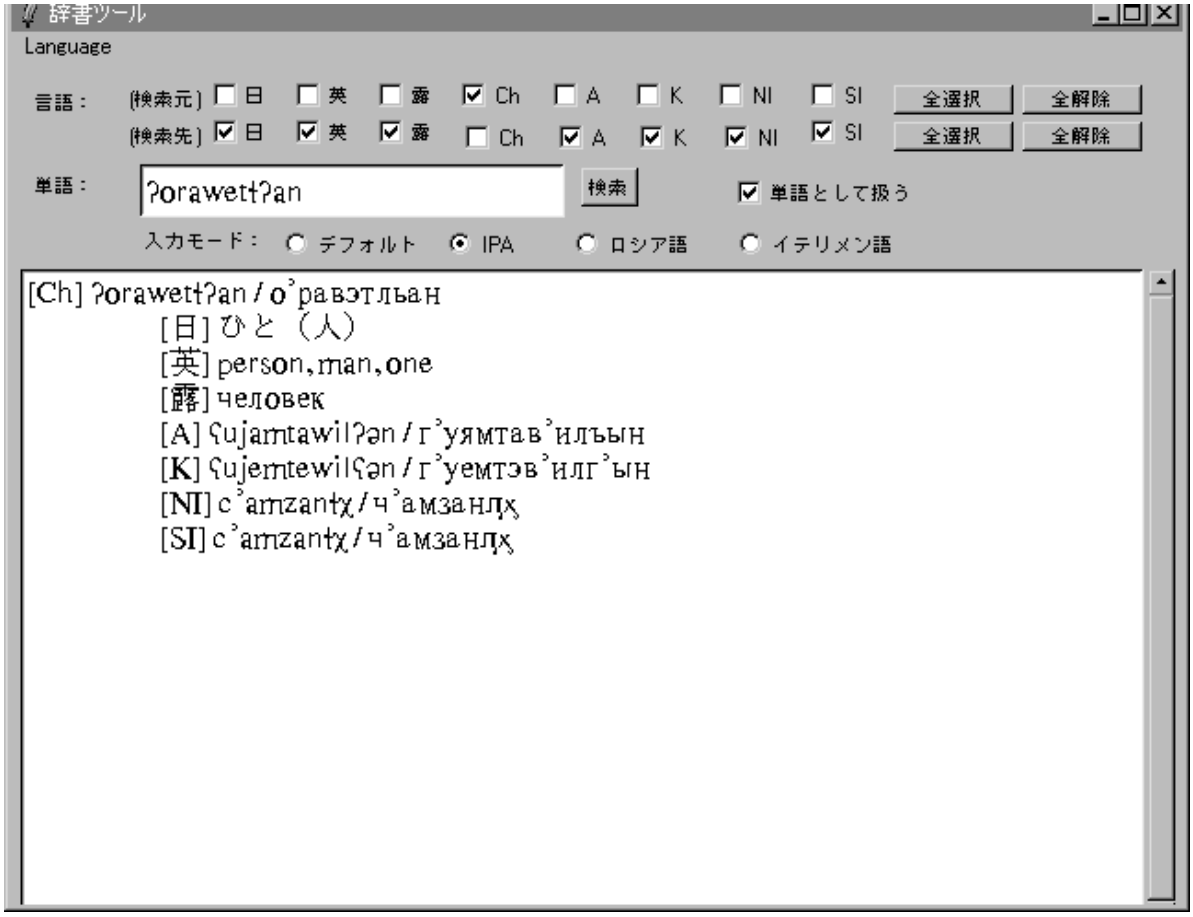
5. トニー・グラハム『Unicode™ 標準入門』翔泳社 2001
6. The Unicode Consortium. *The Unicode Standard. Version 3.0.* Addison-Wesley, 2000.
7. The Unicode Standard, **http://www.unicode.org/unicode/standard/standard.html**.

プログラミング，フォント

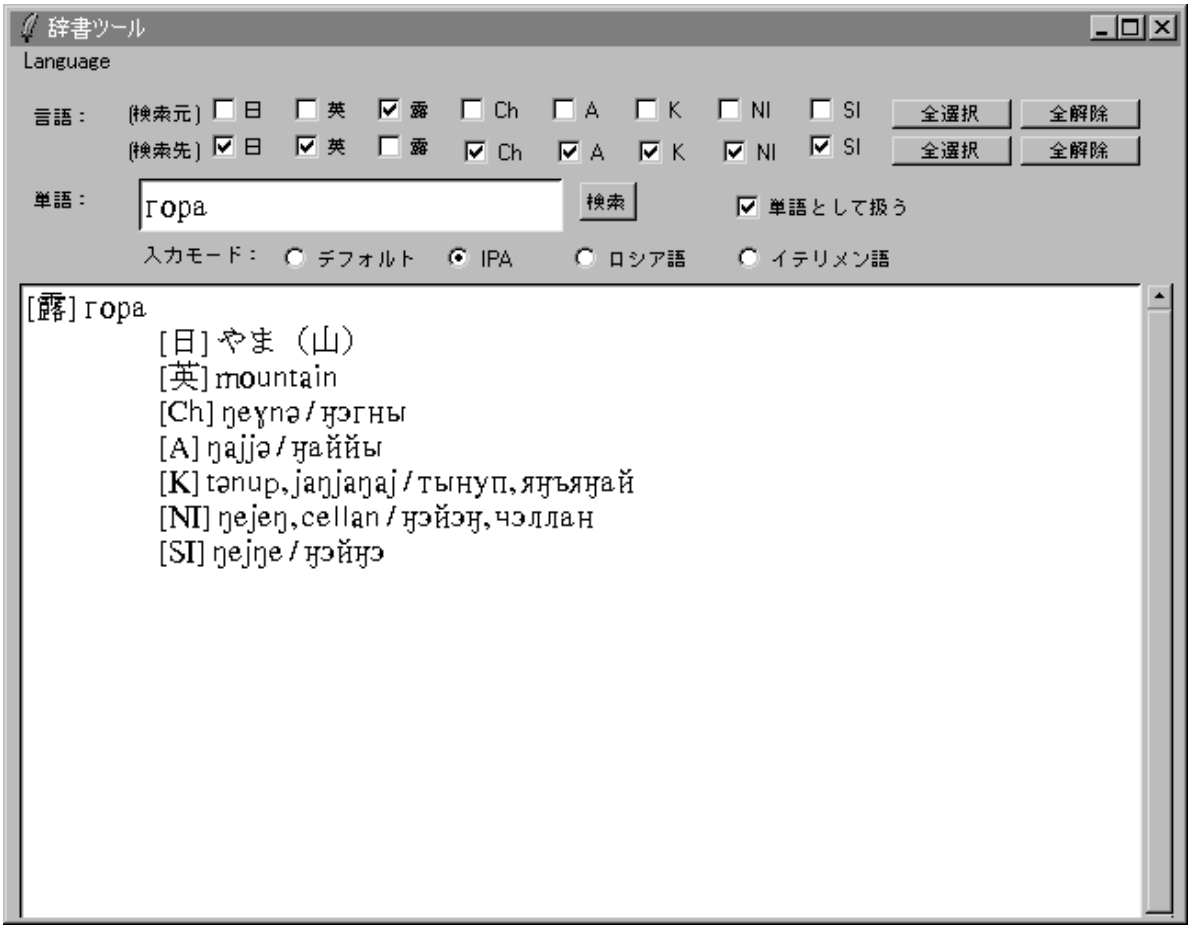
8. 須栗歩人『入門 TCL/Tk』秀和システム，1998.
9. A Little Unicode Editor, **http://mini.net/cgi-bin/wikit/907.html**.
10. Tcl Developer Site, **http://www.tcl.tk/**.
11. The Tcl'ers Wiki, **http://mini.net/tcl/**.
12. Visual TCL project, **http://sourceforge.net/projects/vtcl/**.
13. Free UCS Outline Fonts, **http://www.nongnu.org/freefont**.

本発表は，科学研究費補助金特定領域研究 (2)「消滅の危機に瀕した言語の言語資料のコンピュータ処理のためのデータ構造・分析ツールの研究」(#12039123) による研究成果に基づくものである。

更新日：2002-10-01



【図5】IPA 表記の検索語（チュクチ語 ʔorawetɬʔan）による検索（辞書検索ツール）



【図6】キリル文字表記の検索語（ロシア語 гора）による検索（辞書検索ツール）