京都大学21世紀COEプログラム

東アジア世界の人文情報学研究教育拠点

漢字文化の全き継承と発展のために

国際セミナー　TEI Day in Kyoto 2006
報告書

2006年5月17日(水)

京都大学 百周年時計台記念館

拠点リーダー　高田時雄

(京都大学人文科学研究所)

2006年12月

文字化された言語資源の少ない言語とテクストのマークアップ

松村一登
東京大学大学院・人文社会系研究科・言語動態学専門分野

1. 言語資料

話されたことばは，一過性のものであり，録音したり，文字で書き留めたりしておかない限り，たちまち消えてしまう。書かれたことばでも，メモのように，用事がすむとすぐに破棄されるものが結構あり，出版物は別として，残るものはほんの一部であろう。私たちが一生の間に生み出すことばは膨大なものだが，そのほとんどは，このように文字通り消えてしまい，記録に残らない。

話されたものであるか，書かれたものであるかを問わず，録音，文字表記など，何らかの形で残されたことばの記録のひとつひとつを「言語資料」と呼び，ある言語の言語資料を総体として話題にするときには (X語の)「言語資源」という言い方をすることにする。日本語や英語は，言語資源の豊かな言語であるが，アイヌ語やサーミ語は言語資源が少ない言語である。

定義上，言語資料は文字化されているとはかぎらないが，本稿では，とくに区別の必要がある場合を除いて，通常「言語資料」を「文字化された言語資料」と同義で用いる。文字化された言語資料のうち，機械可読な形態のものを「電子化された言語資料」と呼ぶことにする。〔注１〕

2. 世界の言語的多様性

Ethnologue の名前で知られる Web サイト([14]) のデータによれば，現在，地球上で実際に使われている言語の総数は 6912 であるという (図１)。6900 の言語の母語話者の数はまちまちであり，母語話者人口が１千万人以上の 83 言語 (世界の言語数の 1.2%) の母語話者の合計だけで世界の総人口の８割 (79.46%) を占める一方で，

言語数の上で全体の８割強 (82.1%) を占めている母語話者人口が 10 万人に満たない言語の母語話者数の合計は，世界の総人口のわずか 1.16%にしか達しない。

世界の言語数は，1950 年代には 2000〜3000 ([5])と考えられていたらしいから，単純に数字だけを見て，半世紀の間に言語の数が２倍以上に増えたと受けとめる人がいるかもしれない。しかし，これは，交通機関の発達などにより，世界のすみずみまで調査が進んで，50 年前には知られていなかった言語が多数「発見」されたことによる見かけの増加として理解するのが適当であって，地球上の言語の数は，長期的に見ると，むしろ急激に減少し続けているという見方が支配的である。悲観的な予測によれば，世界の言語は今後，２週間に１言語程度の割合で消滅しつづけ，100 年後には，その数が半分くらいになってしまうとさえ言われている ([7], [8])。

動植物に関して「絶滅寸前種」(endangered species) が語られるように，人間の言語についても，「消滅(の危機)に瀕した言語」(languages in danger of disappearing)，あるいは「危機言語」(endangered languages) という言い方がされる ([16])。〔注２〕「Ｘ語は危機言語か」という質問をしばしば受けるが，理論上はともかく，ある言語が危機言語かどうかの客観的な基準 (例えば，話者人口)があるというよりは，「何らかの理由で１〜２世代のうちに話し手がいなくなってしまうことが危惧される言語」はすべて危機言語だと見なすのが，現場では，いちばん無難な危機言語の定義であると筆者は考えている。

話し手がいなくなって使われなくなる言語がある一方で，新しい言語が誕生しているのもまた事実である。その背景には，「言語」の定義が少しばかり変わってきたことがある。新しく生まれた言語の多くは，これまでどこかの言語の「方言」として扱われてきたものである。言語学者たちは従来，ある言語コミュニティの話す言語形態が，独立した言語であるか，あるいは，隣接する言語形態と「方言」の関係にあるかは，語形の類似の度合いや相互理解の可能性の度合いなど，外から観察できる客観的な要因によって決めることができると考えてきた。しかし，最近は「言語」であるか「方言」であるかを決める際には，言語コミュニティの話者の意

識という主観的な要因が大きな役割を果たすという考え方が主流になりつつある。それゆえ，伝統的には何らかの言語の「方言」とされている言語形態でも，危機言語と呼ぶことがしばしば行われている。

地球上の言語の世界で一種の新陳代謝が行われているだけで，少数言語の盛衰をいちいち気に掛ける必要はないと考える人がいるかもしれない。だからといって，話し手がいなくなってこの地上から消えてゆく言語がまったく記録されないまま消えていってもかまわないということにはならない。危機言語は，使われなくなってしまわないうちに，しかるべき方法で記録にとどめておくべきである。危機言語の言語資料を記録にとどめることを目的とする言語学者たちの研究活動は「言語の記録」(language documentation)，あるいは「記録言語学」(documentary linguistics) などと呼ばれる ([11], [15])。このほか，危機言語を再活性化 (revitalization) させようとする活動も各地で行われている。

3. 文字化された言語資源の少ない言語

数の上で世界の言語の８割をしめる母語話者人口が 10 万人に満たない少数言語のほとんどは，上に述べた危機言語のカテゴリに属すると考えてよいと思われる。これらの言語は，ほとんどの場合，日常的な場面で話す言語として用いられるのみで，文字言語として使われる伝統をもたない，いわゆる「文字のない言語」であると考えられるから，文字化された言語資料がないのは，当たり前と言えば当たり前である。

このような言語すべてが，文字化された言語資料をまったく残さずに消滅してしまう運命にあると言い切るのは言い過ぎである。たとえば，言語学者などが，音声表記 (phonetic transcription)を用いて文字転写し，注釈を加えたり，翻訳を添えて，学術出版物として「X語テクスト集」のようなタイトルをつけて，少部数ながら印刷出版した言語資料が存在する言語も少なからず見られるからである。

「文字をもたない言語」の姿を，音声記号 (phonetic alphabet; cf. [6]) などによっ

て文字転写した原語テクストに，言語学的情報 (形態，品詞，意味など)を付加した言語資料の形で記録し，保存していくことを主要な目的とする研究活動が，言語の記録(language documentation) あるいは記録言語学 (documentary linguistics)という名前で，言語学の新しい分野として認知されるようになったのはここ 10 年くらいのできごとだが，言語学者たちによるこういった研究活動そのものには，音声記号の考え方と同じくらいの歴史的伝統がある。〔注３〕

この種の言語資料は，研究者ごとにさまざまなフォーマットで編集されている。例として４種類のサンプル(図２～５)を掲げる。

図２― ウデヘ語 (ツングース系)，音素表記，日本語訳 ([9] p.104)
図３― アイヌ語，音素表記，日本語訳 ([10] p.364)
図４― ベプス語 (ウラル系，ロシア)，音声表記，ロシア語訳 ([12] p.36-37)
図５― カレリア語 (ウラル系，ロシア)，音声表記，エストニア語訳 ([13] p.175)

ここで，音素表記 (phonemic transcription)と音声表記 (phonetic transcription)の区別は厳格なものではない。アイヌ語の言語資料 (図３)のように，カタカナ表記とローマ字表記が一種の正書法として定着しているケースがある一方で，それとは対照的に，音声記号と複雑な補助記号を多用して，できるかぎり精密な音声転写を行おうとする伝統の行われているウラル系諸語 (図４・図５) のようなケースもある。音素表記(＝表音文字による比較的簡略な表記)と音声表記(＝音声記号を用いたより精密な表記)の違いは，音声表記の精密さの程度の問題といっていい。

言語資料には，原文テクストに，１つ１つの単語ごとに，品詞分類や形態分析などの言語学的な情報を付加し，さらに，１つ１つの文ごとに，英語や日本語のような一般になじみのある言語への翻訳を付加するのが，言語学における標準的な慣行である。筆者が研究対象としている言語の１つ，ウラル系のマリ語の文を使って例示するならば，つぎのようになる。

tošto　jüla-m　kudalt-m-em　ok　šu　[原語]

古い　習慣-対格 廃棄する-分詞-１単　否定-３単 届く　[注解]
「私は古い慣習を捨てたくない」　[翻訳]

原語部分は，この例では，形態素レベルまで形態分析が行われているが，図３・図４・図５のように，単語レベルまでの分析 (分かち書き) しか行っていないこともある。また，注解の形式に一定の決まった方式があるわけではなく，言語ごとに，研究者ごとに，まちまちであるのがふつうである。さらには，図２～図５のように，注解が施されていない言語資料も多い。注解がない場合は，翻訳だけが付加情報として添えられているわけだが，その翻訳も，図２・図３のように，単語ごとの逐語訳，ないしは，行ごとの逐語訳に近い形で与えられることもあれば，図４・図５のように，ただ併記されていたり，対訳になっていたりするだけであることも少なくない。

注解の主体が文法的な情報であることからも明らかなように，この種のテクスト集は，もともとは，言語学者による文法研究のための一次資料として用いられることを目的として編まれたものである。筆者が専門とするウラル系バルト・フィン諸語の少数言語の場合，このようなテクスト集が，20 世紀の中頃から後半にかけて，いくつも出版されているが，これらの言語の話者が激減した現在，この種のテクスト集は，研究者たちだけでなく，当該言語のコミュニティからも，非常に貴重な文化遺産として再評価されるべきものになっている。

4. 少数言語の言語資料の電子化 ― どのような困難があるか

言語資料の電子化の前提は，それが文字転写されていることである。当然のことだが，文字言語の伝統がない言語の場合，文字転写ができる前提は，話されたことばを録音したものがあることである。文字言語の確立している日本語の場合でさえ，談話資料の文字化にはたいへんな労力 (大学院生，若手研究者など，読み書きのできる母語話者を多数動員！) と資金が必要とされることは周知の事実である。文字言語の確立していない言語の録音資料を文字転写（音声表記）する作業のためには，

音声学などの専門の訓練を受けた研究者を養成する必要がある。

本稿は，文字化されている言語資料のマークアップについて論じるのが目的なので，以下では，音声資料の収録とその文字化の作業の段階で遭遇するさまざまな問題についての議論はさけて，すでに文字転写が行われている言語資料があるものとして話をすすめる。音声資料の収録と文字化の作業の中で現れる様々な問題については，別の機会に論じることができれば幸いである。

上で実例をみたような言語資料の提示の方式は，現在，大部分の言語学者にとって，自分自身の研究のために資料を整理し保存するときのフォーマットであると同時に，印刷物として刊行される研究成果の中で言語資料を公開するときの標準的なフォーマットにもなっている。このため，言語資料を，このようなフォーマットでプリンタ出力できるように，ワープロソフトで清書したり，表計算ソフトで整理している言語学者が多いのは当然の成り行きだが，ワープロや表計算ソフトによる文書作成作業，すなわち，印刷用の完全原稿，いわゆる camera-ready の版下を容易に準備できるような書式で言語資料を整理する作業を，言語資料の電子化と同一視している (と思われる) 言語学者が圧倒的に多いように見受けられる。

ワープロの普及で，少数言語のテクストであっても印刷用の版下を簡単に作ることができるようになったことは喜ばしいことである。しかし，その反面，印刷用の版下としての役割を果たした後，有効な再利用の機会のないまま，言語学者のハードディスクの中でワープロ文書のまま眠っている貴重な言語資料も少なくないようである。とくに，ワープロ文書は，見た目に正しい文字として印刷され，たいていの人に満足感を与えてしまう点がくせもので，ラテン文字以外の文字や音声記号などの特殊文字を Unicode で処理することが現在の技術でも十分可能であるにもかかわらず，現状では，大部分の言語学者がまだその利用のしかたを知らないでいる。

言語資料を，他の研究者にあげても文字化けしてしまい，活用してもらえない可能性の高い形でせっせと保存している自分たちが，実はたいへん「もったいない」ことをしているのだと気づく研究者たちが増えてくるのをじっと待ち望まなければ

ならないのは，なんとも歯がゆいかぎりだが，本稿では，言語学者たちが，Unicode や XML，マークアップのような最新の言語技術 (language technology, LT) が，自分たちの仕事において果たす役割に気づくのは，もはや時間の問題だという楽観論に立って，現時点で，言語学者たちがこれらの最新技術を活用しようとした場合にぶつかるいくつかの問題を指摘してみたい。

特殊文字体系や音声記号を，最初から Unicode 文字として入力しようとすると，現状ではまだ様々な技術的制約があり，誰でも簡単に始められるというものではない。実際，筆者の所属する東京大学の言語動態学専門分野では，大学院前期課程に入学してくる学生を対象に，毎年，半年間ではあるが，筆者の過去の科研費で開発したツールやフォント([3][4])を実際に導入しつつ，テクスト処理のための Perl 言語によるプログラミングの初歩，テキストエディタの使い方，音声記号などの特殊文字を Unicode で入力するツールの使い方などの手解きをする実習の授業を開講している。

特殊文字や音声記号の多くは，通常，キーボードから直接入力できないから，録音資料を起こして音声表記したまとまったテクストを入力するためには，特殊文字の入力を容易にする補助ツールが欠かせない。たとえば，Windows XP では，キーボードののキー配列を「入力言語」を選んで設定することで，ある程度カスタマイズできるが，「入力言語」のリストに登録されている言語は，世界中の言語のごく一部にすぎないから，少数言語や特殊な文字セットを用いる言語になればなるほど登録されていない可能性が高い。すべての言語ひとつひとつに文字入力のためのキーボードツールを作るよりは，一般ユーザにもカスタマイズしやすいように設計され，十分な汎用性を持つ特殊文字入力用補助ツールの開発は必須である。

なお，これはマークアップ技術の側の問題ではないが，少数言語の場合，個別の言語名コードが国際規格としてまだ決まっていない ([17][18]) ことがあって，言語資料のマークアップの際に，しばしば戸惑う原因となる。たとえば，ウラル諸語の言語名コードを調べてみると図８のようになっている。つまり，国際規格として認

められた言語名コードが，まだ提案中のままになっているものが，ウラル諸語のほぼ半数に達するという現状は，どうひいき目にみてもあまり感心できるものではない。

少数言語になればなるほど，言語コミュニティーにおける有能な言語の使い手や研究者などの人的資源に限界があり，複数の人間による役割分担による効率化が実現しにくく，ひとりの人間が，現地調査によるテクスト収集，テクストの入力，テクストの整形とマークアップまで，家内工業的に，すべての工程に通じていないとことがスムーズに運ばないことが多い。しかし，言語資料が大量に蓄積されてくるにつれ，実際問題として，何らかの役割分担を導入せざるをえなくなることは目に見えている。そのためには，最新の言語技術に通じた人材を少数言語コミュニティからも育成できるようなしくみを整備していくことが望ましい。大学の言語学系の学科や研究機関が積極的に貢献できるとすれば，まず，このような局面においてでであろう。

筆者のこれまでの研究では，音声記号やキリル文字系の特殊文字で表記された言語資料を Unicode で処理するためのフォントや文字入力ツールの開発や，その利用のためのノウハウの蓄積に重点をおいてきた ([1], [2],[3]) 。たとえば，図５のタイプライタと手書きで書かれた音声表記の部分 (ウラル系カレリア語ジョルジャ方言 [13]) は，ウラル言語学で用いられる国際音声字母とは別の音声記号を Unicode 入力し清書したものが図６である。図４のウラル系ベプス語原文とロシア語訳との対訳資料 (約 290 ページ; [12]) は，図７のような形でタグ付けを施す段階まで作業が進んでいる。

2006 年度から新たに始まった新たな科研費プロジェクトでは，一段階進んで，音声記号や特殊な文字で表記された言語資料のマークアップを，文法研究のような狭い意味での言語学的な研究だけでなく，テクスト分析を行う隣接分野の研究でも実際に活用できる形で行うことに重点を移した研究計画をたてている。個々の少数言語について，マークアップされた言語資料を一定量整備し，それを用いて，文法，

語彙，言語変化などの言語学的研究や，テクスト分析の研究を実際に行うことを目指している関係で，具体的にどのようなマークアップ規格を採用するかを決めるにあたっては，実際にコーパスが実現できる方式かどうかを基準にするという現実路線をとることになるだろうと考えている。TEI はマークアップの有力な候補になっているが，XML でマークアップされたテクストを検索するときの技術的問題が，一般にどの程度乗り越えられるものなのかによっても，筆者の研究プロジェクトの今後の展開はかわってくると思われる。

ちなみに，現時点では，XML によるマークアップの階層性を利用したテクスト検索はまだ最先端の技術であって，人文系の研究者，とくに少数言語を研究対象としている研究者たちが気軽に利用できる段階には至っていないのではないか，というのが筆者の受けている一般的な印象である。しかし，大規模コーパスの老舗である BNC コーパス (British National Corpus) が XML マークアップに移行しつつあることを考えると，XML を生かしたコーパス検索ツールが本格的に普及する日は案外早いかもしれない。

5. 最新の言語技術がなぜ少数言語に重要なのか

近い将来，文字言語資料のほとんどが，コンピュータを使って作成された電子的な文書となり，そのままコンピュータを媒介にして交換され，かつ，コンピュータやネットワークに蓄積されて情報検索に利用される時代が到来するものと思われる。そのとき，WWW はもちろんのこと，最新の言語技術 (LT) の恩恵に浴すことのできない少数言語は，残念ながら，現状のままでは，その使用領域が加速度的に縮小していく運命にあるだろうと懸念される。使用人口が少ない言語になればなるほど，言語使用の現場において，電子的に再利用可能な言語資料を産みだすための技術的な制約がますます増大し，その結果，その言語の使用の機会がさらに減少するという悪循環が予想されるからである。もし現実にそうなれば，少数言語を母語とする人々は，これまで以上に，母語のかわりに，多数派の言語で生活することを余儀な

くされ，強力な大言語への同化がますます促進されることになる。しかし，一昔前ならいざ知らず，大言語への同化が進むことを歓迎すべきこととして，成り行きに任せるのをよしとする大言語中心の立場をとり続けるわけにはいかない。

誰も，自分が特定の言語を母語としていることにより，日常生活において不利な状況に置かれることがないように配慮された社会を，言語的にバリアフリーな社会と呼ぶことにしたい。少数言語を母語とする言語的少数者が暮らしやすい社会の方が好ましいと考えるなら，少数言語が社会の隅に追いやられ，次第に消滅していくのではなく，今後も安定して使われ続けることが可能になる条件を社会的に整えて，少数言語を母語とする言語コミュニティを支えてゆく必要がある。

少数言語が今後も健全な形で生き残ることができる前提は，少数言語の立場を考慮した言語技術 (language technology) が確立されることである。言語資料に乏しい少数言語コミュニティでは，現在，新聞発行，出版活動などにおいて，もともと大言語の使用者のニーズに応えるために開発された商業的なワープロソフトやＤＴＰソフトを苦労して使って，自分たちの言語による印刷物を発行するための版下作りを行っている。これでは，他の用途への転用や再活用の可能性が非常に限定された言語資料を次々と産出するために，少数言語の有能な使い手たちの能力が浪費されているといっても言い過ぎではない。少数言語の言語資料も，大言語と同じように，最先端の言語技術を駆使したコンピュータ処理ができるような環境を整えて行くことが望まれる。

6. 展望 — 電子文献学，もしくは，言語資料学

こういった方向に社会が進んで行くためには，伝統的にことばと深く関わってきた大学の文学系・外国語系の学科や，人文社会系の研究機関が，最新の言語技術を使った言語資料の産出・処理のできる人材を育成するための研究や教育に，本格的に取り組む必要があるだろう。この点では，京都大学の人文科学研究所をはじめ，東京外国語大学のアジア・アフリカ言語文化研究所，あるいは大阪の国立民族学博

物館のように，人文系の研究者と情報処理系の研究者が共同研究を組みやすい研究機関の方が，大学の文学系・外国語系の学科と比べ，すでに一歩先んじているかも知れない。

言語の研究は，ヨーロッパの伝統では，もともと philology と呼ばれていたことはよく知られている。Philology を「文献学」と訳したのは，京都大学教授をつとめた詩人の上田敏 (1874-1916) らしいが，「文献学」とならんで「博言学」という訳語も一時期用いられたようである。同じ言語の研究でも，サイエンス，それも自然科学を自称し，「言語学」と訳される 20 世紀の新参者の linguistics と比べると，「文献学」には，図書館の隅の目立たない場所で，誰も関心を持たないような，表紙が変色し，埃をかぶった古い文書を一人孤独にひもといている文系研究者という，あまりあか抜けないイメージがどことなくつきまとう。

広辞苑によれば，「文献学」とは「文献の原典批判・解釈・成立史・出典研究を行う学問」であり，また「それに基づき民族や時代の文化を研究する学問」とされる。よくよく考えてみれば，広辞苑のこの定義は，「文献」を「言語資料」で置き換えれば，電子化された言語資料を出発点として展開されようとする言語の研究のありかたそのままである。とすれば，近頃聞かれる「電子文献学」(computational philology) という形で，言語資料の電子化やそのマークアップに取り組む人文系の研究分野が，この由緒正しい「文献学」という名前を継承するのは理にかなっているように思われる。

ここで注意してほしいのは，「電子・文献学」であって「電子文献・学」ではないことである。従来の「文献学」は，対象が，手稿や印刷物の原典の場合でも，マイクロフィルムなどによる原典の複写されたものである場合でも，人間の目を道具としてテクストを読み，情報を抽出するという方法で行われてきた。「電子・文献学」とは，従来人間の目が担ってきた役割の少なくとも一部をコンピュータに担わせる形で行う「文献学」である。いいかえると，情報抽出やその分析の方法が電子的であるという意味で「電子・文献学」なのである。これに対して，「電子文献・

学」は，従来型の紙のメディア上の文献と対立する意味での電子メディア上の文献，すなわち「電子的な文献」「デジタルな文献」を研究対象とする分野の意味であって，この場合，電子的であるのは研究対象の方である。

出版形態が急速にデジタル・メディアの方向に移行しつつある現在，新しく作られていく文献については，「電子・文献学」でも「電子文献・学」でも，事実上ほとんど違いは出ないかも知れない。しかし，電子メディアによる文献製作が登場する以前の文献を対象とする歴史学や古典学のような従来型の人文系研究分野からみると，「電子・文献学」と「電子文献・学」の間には，天と地ほどの開きがある。電子化された言語資料をマークアップすることで，コンピュータによる文献からの情報抽出を容易にしようとする方向は，まさに「電子・文献学」と呼ばれるにふさわしい。

電子的な言語資料をコンピュータを使って処理することが目新しかった時代ならともかく，それが普通になりつつある時代に，「電子文献学」などと，わざわざ「電子」を冠して呼ぶのはわずらわしいから，単純に「文献学」でいいではないか，という考え方もあるかもしれない。それも一理ある。「文献学」は，たしかに器は古いかも知れないが，新しい酒を盛ってはいけないという決まりはない。今のうちは「電子」とつけて目新しさを誇示してはいるが，いずれは「電子」がとれて「文献学」という名前になる可能性も大いにある。

他方，筆者は，「言語資料学」という名称を数年前から使っている。外国語教育や辞書編纂の基礎と見られがちで，用途が限定されている (言語学的な) コーパスはもちろん，歴史学，民俗学などなど，「ことば」に関わる分野で利用されているテクスト系の言語資料をすべてカバーする概念として「言語資料」(linguistic document) という考え方を前面に出そうという意図である。今後，言語学の下位分野としてのコーパス言語学や談話研究などは，文法研究・語彙研究といった限られた用途のために特別に構築された，狭い意味でのコーパスを使うことから，たとえば，国会の議事録のような言語資料をも，ふつうに研究の対象とするように拡張されていくに

違いない。とすれば，コーパス言語学は，今後，文学研究，歴史学，フォークロア研究，ライフ・ヒストリー研究といった，文字化された言語資料から情報を抽出することが研究の出発点ないし重要な部分になっている人文系の研究分野との共通性をしだいに高めていくことになると思われる。このように考えれば，言語資料を拠り所として展開される研究諸分野を総称する概念として，「言語資料学」を用いてもかまわないのではなかろうか。

名称はともかくとして，言語学の本流の方が，いつのまにか文法研究さえも通り過ぎて，脳科学の方向にどんどん傾斜し，言語資料からどんどん離れている感のある今，言語の研究にとっての原点である言語資料を，ふたたびクールな研究対象として復権することができそうな展望が開けてきたことは，とても喜ばしいことである。テクストのマークアップは，そのためのキーワードのひとつである。

注

〔1〕文法研究や辞書編纂などを目的とする場合，テクストを検索して，ＫＷＩＣ索引などの形で用例を探したり，特定の語についての使用頻度などを調べる必要がある。残念ながら，現在の技術では，デジタル録音された言語資料を直接，自由自在に検索することができないので，何らかの方法で文字転写したものを代わりに用いている。言語の記録という観点からは，文字転写したものは，一次資料としての音声データに対して補助的な役割を果たすにすぎない訳だが，文法研究や辞書編纂の現場では，この関係が逆転し，文字転写し，マークアップしたものの方を主たる言語資料として扱い，一次資料である音声データの方を補助的なものと見ているのが現状である。

〔2〕「危機言語」は，1990 年代の中頃，日本言語学会で endangered languages の日本語訳として採用されたもので，今ではすっかり定着した。ちなみに，中国語では，これを「瀕危語言」と呼ぶようである。

〔3〕音声記号の国際的な標準として知られている「国際音声字母」(International Phonetic

Alphabet, IPA) を提唱した国際音声学会 (International Phonetic Association) の設立は 1886 年。エジソンによるロウ管式蓄音器の発明は 1877 年で，19 世紀末から 20 世紀の初頭には，ロウ管による「珍しい言語」の音声の録音が盛んに行われている。文字を持たない少数言語のテクストが文字化されて，言語学の出版物の中にふつうに現れるようになるのは，おおむね 1910 年代以後である。

参考文献・Web サイト

[1] 松村一登 2002「ロシアのウラル諸語の言語データの収集とその電子化の試み」—『危機言語の現地調査および記述的研究』(平成 11〜12 年度科学研究費補助金基盤研究 (A)・研究成果報告書), pp.51-68

[2] 松村一登 2006「マリ語の言語資料とその電子化」Uralica, Vol. 14 (2006) [印刷中]

[3] 松村一登 2006「ウラル系諸語の言語資料の電子化とマークアップ」—『音声記号等で表記された言語資料のマークアップとコンピュータ処理』(平成 15〜17 年度科学研究費補助金基盤研究 (A)・研究成果報告書), pp.1-30

[4] 鈴木麗爾，小野智香子，松村一登 2003 『フィールド言語学者のための Unicode ツール』(環太平洋の「消滅に瀕した言語」にかんする緊急調査研究成果報告書 B010)

[5] 市河三喜・服部四郎 (共編) 2006『世界言語概説・下巻』研究社

[6] 『国際音声記号ハンドブック』 大修館書店, 2003

[7] デイヴィド・クリスタル 2004『消滅する言語 — 人類の知的遺産をいかに守るか』中公新書

[8] ダニエル・ネトル，スザンヌ・ロメイン 2001『消えゆく言語たち — 失われることば，失われる世界』新曜社

[9] 風間伸次郎 2006『ウデヘ語テクスト２』東京外国語大学アジア・アフリカ言語文化研究所

[10] 静内町文化財調査報告 1995『静内地方の伝承Ⅴ—織田ステノの口承文芸(5)—』静内町郷土史研究会

[11] Nikolaus P. Himmelmann 1998. “Documentary and descriptive linguistics,” in *Linguistics*, Vol.36, No.1: 161-95.

[12] М. Зайцева и М. Муллонен 1969. *Образцы вепсской речи*. Издательство “Наука”, Ленинградское отделение.

[13] Jaan Õispuu 1990. *Djordža karjala tekstid.* Tallinna Pedagoogiline Instituut.

[14] http://www.ethnologue.com (Ethnologue. Languages of the World)

[15] http://www.hrelp.org/documentation (Language Documentation, SOAS)

[16] http://www.tooyoo.l.u-tokyo.ac.jp/ichel/ichel-j.html (危機言語のホームページ)

[17] http://www.loc.gov/standards/iso639-2/langhome.html (Codes for the Representation of Names of Languages)

[18] http://www.sil.org/iso639-3/ (ISO 639 Code Tables, SIL)

図１　母語話者数を基準にした世界の言語の分布 ([14] による)

母語話者数のクラス	現在使われている言語			母語話者数		
	実数	%	累計(%)	実数	%	累計(%)
100,000,000 ～ 999,999,999	8	0.1	0.1	2,301,423,372	40.21	40.21
10,000,000 ～ 99,999,999	75	1.1	1.2	2,246,597,929	39.25	79.46
1,000,000 ～ 9,999,999	264	3.8	5.0	825,681,046	14.43	93.88
100,000 ～ 999,999	892	12.9	17.9	283,651,418	4.96	98.84
10,000 ～ 99,999	1,779	25.7	43.7	58,442,338	1.02	99.86
1,000 ～ 9,999	1,967	28.5	72.1	7,594,224	0.13	99.99
100 ～ 999	1,071	15.5	87.6	457,022	0.01	100.00
10　～ 99	344	5.0	92.6	13,163	0.00	100.00
1 ～ 9	204	3.0	95.5	698	0.00	100.00
不明	308	4.5	100.0			
計	6,912	100.0		5,723,861,210	100.00	

出典: http://www.ethnologue.com/ethno_docs/distribution.asp?by=size#2 (as of 05-05-2006)

図２　ウデヘ語の言語資料（[9] p.104）

2005年3月14日 クラースヌイ・ヤール村にて録音
N. P. Kukchenko 氏 口述

6. bii guufui-də sagdi samaa bisə
私の 叔父は 偉大な シャーマン だった

77−001
gə xaisi omo bii guufui-də xaisi sagdi samaa bisə.
さあ、もう 一人、私の 叔父（父の姉妹の夫）も やはり 偉大な シャーマン だった。

77−002
guufusini, *znaesh'* sagdi samaa, səwəsiləmi.
亡くなった叔父は、わかるか、偉大な シャーマンで、シャーマンをすると、

77−003
səwəsiləmi sikiə, dəŋə təu puundəgiwənəini bisə.
シャーマンすると、夕方、灯りを 全部 消させるの だった。

77−004
''nii-də əǰiu saulagi. bii jəu-də waami təxəsiǰəŋəi.
「誰も 火を点けるな。私が 何の獣の生贄でも 殺して 皮を剥ぐだろう。

77−005
təxəsiək woosiǰaŋai. təu woosii mətəisini saulagitəuǰə,
剥いで あちこちに投げるだろう。全部 投げ 終わったら 火を点けてくれ、」と、

77−006
gə utə, buji waa-bədə, bujiwə, təxəsii-bədə, təxəsii bisə,
さあ そうして、獣を 殺しているようだ、獣を、皮を剥いでいるようだ、剥ぐの だった、

77−007
uti, dogbo, səwəsimi. mətəə-tənə dianaini, ''gəə, saulagija-ǰa,
彼は、夜に、シャーマンして。終わると、言う、「さあ、灯りを点けろ、」と。

77−008
joktosiami jəu-də anči. təu gaagisii, jauxi, jauxi-ka gaagisiini.
私はよく見たが 何も 無い。全部 持ち去った、どこへ、どこかへ 運び去った。

77−009
tuə-tənə, wakcami, xulimi, baazagatigini tuə, sualaǰami.
冬、 狩りして、歩き回って、タイガへ 冬、スキーで行く。

77−010
ono sualaǰami bimi ono sualai ujələni,
どうやって スキーで行って いるのか、どうやってか 自分のスキーの 上に、

77−011
kəptəmi ŋuai toowa-da əi ila,
横になって 寝る、火さえも 焚かない、

図３　アイヌ語の言語資料（[10] p.364）

	ペッ トゥラシ ウテレケレアン	
	pet turasi uterkere=an	川をさかのぼって走って行った。
	レラメトク アンネプ ネクス	
	re rametok an=ne p ne kusu	私たちは三人の勇者なので
	ホシキノ テレケアンコンノ	
	hoski no terke=an konno	先に私が走ると
	イオシ アコロ オッカイポウタラ	
2040	i=os a=kor okkaypo utar	私のあとから私の若い者たちも
	イオシ	
	i=os	私のあとから
	イノシパ フミ ネコトム イラムアン	
	i=nospa humi ne kotom iramu=an wen …	追いかけてくる音がするように思えた。
	ウェン…ケウトゥム ウェンカ アキ	
	wen … kewtumu wen ka a=ki	私は思いが悪くもなり
	ウェン イルシカ アキ	
	wen iruska a=ki	ひどく怒りもした。
	ウェンカムイウタラ エネ ヘンネ キヤクン	
2045	wenkamuy utar ene henne ki yakun	悪者たちがあんなふうにしなかったら
	アンロンノカ ソモ キ	
	an=ronno ka somo ki	私は殺しもしない
	アンチセウコウフイカカ ソモ キイケ	
	an=ciseukouhuyka ka somo ki hike	家といっしょに燃やしもしなかったのに
	ウェイサンペ コロワ	
	wen sampe kor wa	悪い心がけを持って
	アコロ オナ コロペ オピッタ ルラクス	
	a=kor ona kor pe opitta rura kusu	父さんのものを全部持ち去ったので
	ウェンカムイウタラ ポクナモシリ アコキル	
2050	wenkamuy utar poknamosir a=kokiru	悪者どもを地獄に突き落としたのだ。
	アコシユプ カムイ ヌプリ	
	a=kosiyupu kamuy nupuri	私は気合いを入れて神の山
	ヌプリ トゥラシ スイ	
	nupuri turasi suy	山を登ってまた
	テレケアニネ リキプアニネ	
	terke=an hine rikip=an hine	走って登って
	アコロ オナ コロペウタラ	
	a=kor ona kor pe utar	父さんの持っていたものを
	スイ アンウサライェイネ	
2055	suy an=usaraye hine	また分け合って
	アイセ エアシカイ パクノ	
	an=se easkay pakno	運べるだけ

図４　ベプス語の言語資料（[12] p.3637）

Иняков Василий, 15 л.
Маг. зап. 166/3. М. Муллонен, 1961.

16. príha ostí bazarou źerklon

elí endę ükś príha. ťedan, mäń hän bazaralę i ostí śigä źerklon. kodhe tuĺ, kaclęb źerklohо: «čoma olęn». ak ńecęn homęič, dumę,ib: «minak hän śinna kaclęb?». konz hän ĺäkś (príha) kodišpęi, ak kacob: «a, sanob, ńügüde ťedan, keda kaclęb. hän, sanob, bazarou ĺüuźi ńeičen da sęn karťinąine om. tulęške, sanob, mamoi, kaco». mamazę tuĺ, kacuhťi: «ka, sanob, om babka ńečiť, minun vuiťťe». śit tuĺi tatazę: «ka min ťii paǵižętęi, om ńečiť mužik, bardanke da furaškoiš». śid vähäinę da tuĺ iče hän príha. he ǩüzęltasęi: «kedak śina bazarou ĺüuźid?». — «a kedak om?». — «ka mii naku kacuimęi, sanotas, da em ťea, ǩen om». a hän śiižuťi heiť kaikid ühtęs da ozuťi: «vot, sanob, ťii iče olętęǵi».

16. Парень купил на базаре зеркало

Жил раньше один парень. Пошел он на базар и купил там зеркало. Пришел домой, посматривает в зеркало: «Красивый я». Жена заметила это, думает: «Что он туда посматривает?». Когда он ушел из дому, жена смотрит: «А, говорит, теперь знаю, кого рассматривает. Он, говорит, нашел на базаре девушку, да это ее карточка. Иди-ка, говорит, мама, посмотри». Ее мать пришла, посмотрела: «Да это, говорит, бабка, похожая на меня». Потом пришел ее отец: «Да что вы говорите, это мужик, с бородой да в фуражке». Потом через некоторое время пришел сам парень. Они спрашивают: «Кого ты на базаре нашел?». — «А кого?». — «Да мы тут смотрели, говорят, да не знаем, кто». А он поставил их всех вместе и показал: «Вот, говорит, сами вы и есть».

図５　カレリア語の言語資料（[13] p.175）

98. / s/l e p u v u t t i h š i l m ä t /

s/lepuvuttih //

/ k o ž ? /

jo ka počti• vuaž hänel // vuaž // üks silmäst čirkzen nägöw / a toin / toin ei näw // i / i apera•cid ei ruvet ruadmah // näil oll/ah omat / m o s k u š //i to ei ottuče // mož fofse• rikot // razv čirkzen nägüw // ain šanow // "kuin vihmuw / fos-sendah nakroičet / vrod on sol/nj/škan" // tüt ei niä // ei // ühel šilmäl // üksin el i / hänen kodin on atale•nnest kaik küläš // a heidäh i / i kaik siel / on seičmen kodi / i koih müt-ten oška / šin vanhat // nu händ ei kačot / što hiän on / al/e/loš // šid mändih / a hiän kua/liä //

ML Sem 85

/ Silmad jäid pimedaks? /

Jäid pimedaks.

/ Millal? /

Juba peaaegu aasta tagasi. Aasta. Ühe silmaga natuke näeb, aga teine, teine ei näe. Ja, ja operatsiooni ei hakata tegema. Neil on sugulased Moskvas ja needki ei võta (operatsioonile). Võib-olla rikud täitsa ära. Siiski natuke näeb. Ütleb alati: "Justkui sajaks, justnagu oleks päike." Tüdruk ei näe. Ei, on ühe silmaga. Üksi elas ja ta majake on külast eemal. Aga neid on... Seal on seitse maja ja kõik (elanikud on) vanad. Tema järele ei vaadatud, on ta elus või ei. Siis mindi, aga ta on surnud.

図６

ウラル音声表記 (UPA)のテクスト(図５)を Unicode 対応フォントで清書したもの

/ sлepuvuttj̮h šil´mät /

sлepuvuttj̮h //

jo ka počti· vuaž hänel̄´ // vuaž // ükś šil´mäšt čirkzen ńägöw / a toiń / toiń ei ńäw // i / i apera·cid ei ruvet ruadmah // ńäil̀´ oллah omat / moskuš // i to ei ottuče // mōž fofśe·rikot // ràzv čirkzen ńägüw // ain šanow // « kuin vihmuw / fośśendah nakroičet / vròd́ on soлnj̮škań » // t́üt̀´ ei ńiä // ei // ühel̄´ šil´mäl̄´ // ükśin el´j̮ / hänen kod́in on atal´e·nneśt´ kaik̀ kül´äš // a heidäh i / i kaik̀ šiäl̀´ / on seičmen kod́ī / i koih müt´t´eń oška / šin vanhat // nu händ ei kačot /što hiän on / al´ eлoš // šid mändih / a hiän kuaлiä //

図 7

ウラル音声記号で転写されたベプス語のテクスト(図４)に最小限のタグ付けを試みたもの

```
<div3 id="16">
<informant>Иняков Василий, 15 л.</informant>
<doc_info>Маг. зап. 166/3. М. Муллонен, 1961.</doc_info>
<div4 id="16_1" lang="vep">
<h>16. príha ostʹi bazarou źerклon</h>
<p>
<s no="0671"> eĺi endę ükś príha. </s>
<s no="0672"> tʹedan, mäń hän bazaralę i ostʹi śigä źerклon. </s>
<s no="0673"> koďhe tuĺ, kaclęb źerклoho: «čoma olęn». </s>
<s no="0674"> ak ńecęn homęič, dumę,ib: «minak hän śinna kaclęb?». </s>
<s no="0675"> konz hän ĺäkś (príha) koďišpęi, ak kacob: «a, sanob, ńügüde tʹedan, keda kaclęb. </s>
<s no="0676"> hän, sanob, bazarou ĺüuźi ńeičen da sęn kartʹinąine om. </s>
<s no="0677"> tulęške, sanob, mamoi, kaco». </s>
<s no="0678"> mamazę tuĺ, kacuhtʹi: «ka, sanob, om babka ńečitʹ, minun vuitʹtʹe». </s>
<s no="0679"> śittuĺi tatazę: «ka min tʹii paǵižętęi, om ńečitʹ mužįk, bardaŋke da furaškoiš». </s>
<s no="0680"> śid vähäine da tuĺ iče hän príha. </s>
<s no="0681"> he ĺküzęltasęi: «kedak śina bazarou ĺüuźid?». </s>
<s no="0682"> – «a kedak om?». </s>
<s no="0683"> – «ka mii naku kacuimęi, sanotas, da em tʹea, ĺken om». </s>
<s no="0684"> a hän śiižutʹi heitʹ kaikid ühtęs da ozutʹi: «vot, sanob, tʹii iče olętęǵi».
</s>
</p>
</div4>
</div3>
```

図８　ウラル諸語の言語名コード ([17][18] による)

		ISO 639-2 & 639-1	ISO/DIS 639-3
フィン・ウゴル語(その他の)	Finno-Ugrian (Other)	fiu	
エストニア語	Estonian	est (et)	est
フィンランド語	Finnish	fin (fi)	fin
メアンキエリ語	Meänkieli		fit*
カレリア語	Karelian	krl	krl
オロネツ語	Livvi [Olonetsian]		olo*
ベプス語	Vepsian		vep*
イジョール語	Ingrian [Izhorian]		izh*
ボート語	Votic [Votian]	vot	vot
リーブ語	Liv [Livonian]		liv*
サーミ語(その他の)	Sami languages (Other)	smi	
イナリ・サーミ語	Inari Sami	smn	smn
ルレ・サーミ語	Lule Sami	smj	smj
北サーミ語	Northern Sami	sme (se)	sme
スコルト・サーミ語	Skolt Sami	sms	sms
南サーミ語	Southern Sami	sma	sma
アカラ・サーミ語	Akkala Sami		sia*
ケミ・サーミ語	Kemi Sami		sjk*
キルディン・サーミ語	Kildin Sami		sjd*
ピーテ・サーミ語	Pite Sami		sje*
テル・サーミ語	Ter Sami		sjt*
ウーメ・サーミ語	Ume Sami		sju*
マリ語	Mari	chm	chm
東マリ語	Eastern Mari		mhr*
西マリ語	Western Mari		mrj*
エルジャ語	Erzya	myv	myv
モクシャ語	Moksha	mdf	mdf
コミ語	Komi	kom (kv)	kom
ウドムルト語	Udmurt	udm	udm
ハンガリー語	Hungarian	hun (hu)	hun
ハンティ語	Khanty		kca*
マンシ語	Mansi		mns*
セリクプ語	Selkup	sel	sel
ネネツ語	Nenets		nen*
ガナサン語	Nganasan		nio*

[注] * はまだ国際規格として認められていない(提案中)のもの