Ethnologue15th 言語属性データと言語系統データの生成および言語同定における利用

呉靭・乾秀行・杉井学・松野浩嗣

◎Key Words　**語族，言語系統樹，言語同定，HTML タグ，*Ethnologue***

連絡先：山口大学大学院理工学研究科ネットワーク科学研究室
Contact to：wu@ib.sci.yamaguchi-u.ac.jp

1　はじめに

近年，GIS (地理情報システム)を言語学に応用する研究が，文理融合の新たな手法によって言語学的新発見をもたらすことが期待されている。我々はそのような研究を展開するため，GIS データの生成処理[1]から始まり，そこで必要となる「言語同定」の手法[2]を提案した。その手法では，我々は言語を表わす言語名等の属性情報に加え，言語の系統分類も考慮し，世界諸言語に関する言語属性データおよび言語系統データを用いている。

我々は言語属性データとして SilGIS-Data[1][2]を利用するが，その言語名の多数に文字化けが発生している(原因については後述)。さらに言語系統データが含まれていない。一方，*Ethnologue15th*[3]の Web サイト[4](以降「Web サイト」と省略する)には我々が必要としている言語の属性および系統情報が掲載されており，参照できるようになっている。本研究では Web サイトからの情報取得およびデータ生成法を提案する。

以下，2 節では言語の属性と系統のデータついて説明する。3 節では Web サイトの Web ページの構成上の特徴と HTML ソースの解析について述べた後，Web サイトから言語の属性と系統情報を自動取得する手法について述べる。4 節ではデータの正当性を確認するためのチェック処理や変換方法，またその言語同定における利用について述べる。最後の 5 節で本稿をまとめる。

2　言語の属性と系統および言語同定

1 つの言語には，複数の名前が付いている場合がよくある。例えば「日本語」を/nippon-go/と/nihon-go/と二通りに読んだり，あるいは英語読みで「Japanese」と言うようなものである。*Ethnologue15th*では，その複数の言語名の中の 1 つを第一言語名とし，その他は別名とする[3]。言語に関する情報は今挙げた第一言語名や別名の他，話者人口，言語の話されている地域，方言名，言語使用状況など多くあり，それらは言語の属性情報となる。

第一言語名および別名の指定は学者独自に行われているため，異なる学者によって編成された言語データでは，同じ言語が違う言語名 (第一言語名)になっているケースも多い。我々が GIS データの生成処理[1]に利用した言語属性データ SilGIS-Data と言語別語順データ Yamamoto-Data[1][5]についても同様で，ゆえに我々は 2 つのデータの照合および確認を行う必要がある。この処理を「言語同定」[2]と呼ぶことにする。

SilGIS-Data には第一言語名や別名など約 60 項目が含まれている。SilGIS-Data はそのファイル形式の制限により，項目によっては文字列が一部区切られたような不完全なものや，言語名文字列に文字化けが発生している箇所が多数見られる[6]。第一言語名や別名等は言語を同定するための重要な属性情報である[2]ため，文字化けしていない完全なデータが必要である。

世界諸言語は系統的に分類される。SilGIS-Data が準拠している *Ethnologue15th* では世界諸言語を 108 の語族 (Language Family)[7]に分類している[8]。同じ語族に属する言語は，はるか過去に話されていた 1 つの言語から分かれて発展してきたと言語学者は考えている。語族は木構造[9]をなしており(以降，「言語系統樹(Language Family Tree)」と呼ぶ)，そのイメージを Fig. 1 に示す。木構造の特徴から，各々の言語系統樹の葉(leaf)[9]は「言語」，葉以外の節点 (node)[9] は便宜上すべて「言語グループ」とそれぞれ呼ぶことにする。我々の提案した言語同定手法[2] は，例えば Fig. 1 の(A)にある「M」と(B)にある「M̀」を同じ言語と認定できるようにするためのもので，Fig. 1(B)に示しているような SilGIS-Data に関

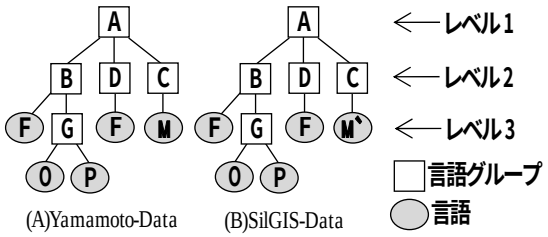

Fig. 1 **言語系統樹**(Language Family Tree)

連する言語系統データが必要となる[5] 。

3 データの生成

3.1 Web サイトの構成およびページリンクの特徴

Web サイトにある言語名別目次ページ (http://www.ethnologue.com/language_index.asp) には A から Z までのアルファベットがあり，そのアルファベットで始まる言語名のページにリンクされている。例えば，「J」をクリックすると「Japanese」などの「J」で始まる言語名リストのページ(以降「2 段階目の目次ページ」と呼ぶ)が表示される。2 段階目の目次ページにある各々の言語名にはその言語に関する属性情報のページ(以降「言語ページ」と呼ぶ)がリンクされている。

各言語ページはその言語が話されている国別(1 カ国以上，数は不定)に情報が配置されていて，第一国(言語の発祥地または最も多くの話者がいるとされている国)[3]の情報は必ず掲載されている。項目としては別名や方言名，分類などがあり，国別の項目数は不定である。第一国においては，系統的分類は言語の所属する言語系統樹における根(root)[9]からその言語の親(parent) [9]までの経路(path)[9]を示しており，必ず表示される項目である。別名および方言名は言語別，国別の情報で，すべての国について必ずしも存在している項目ではない。

語族別目次ページ(http://www.ethnologue.com/family_index.asp)には，すべての語族の語族名がアルファベット順に配置されていて，各々の語族名にはその語族の言語系統樹を表したページ (以降「語族ページ」と呼ぶ) がリンクされている。語族ページにある言語系統樹の節点は，(i) 言語グループ，(ii) 言語，という 2 種類に分けて区別できるように表現されている。その節点のテキストの形式および特徴を Table 1 に示す。それにより節点のタイプが言語グループか，それとも言語かの識別ができる。

Fig. 1 からもわかるように，言語グループは本来必ず子(child) [9]をもつ。しかし，語族ページは横スクロールしないように画面設計されているため，各々の語族ページにおいて，言語系統樹の木が高い[9]場合には，言語グループの下位階層の情報はさらにリンクしている下位の語族ページに含まれていることがある。ある言語グループが見掛け上葉になっている場合は，必ずリンク先が設定されているため，リンクを辿っていけば，各々の語族の全構成要素を表示することができる。

3.2 Web ページの HTML ソース解析

Web サイトの関連ページの URL にいずれも「.asp」が含まれ，また各々のページが HTML[10]形式であることから，このサイトの作成には Microsoft 社の ASP 技術[11]が使われていることがわかる。よって，ページの情報の表示形式やリンク先の設定などに規則性があると考え，ページのソースを解析したところ，ページの構成および HTML タグ[10]の使用に関し次の特徴があることが明らかになった。

(1) 言語別目次ページと言語ページの特徴

A) 言語名別目次ページにある A から Z までのアルファベットのリンク先 URL は「･･･/language_index.asp？letter=X」の形式になっていて，1 桁英字「X」が可変で，その文字である。2 段階目の目次ページも類似の構造となっているため，説明は省略する。

B) 言語ページにある言語名 A，第一国の国名 B，言語コード zzz はそれぞれ「<H1> A </H1>」，「<h2>A language of <A HREF="･･･">B</A></h2>」，「 ISO639-3: zzz 」の文字列を分割することにより取得できる。

C) 第一国における属性情報は<TABLE></TABLE> (表)にまとめられている。1 行(<TR>･･･</TR>要素)が 1 項目の情報を表わしていて，左列 (<TD>･･･</TD>要素)が項目名(例として，別名:「Alternate names」) で，右列がその内容となっている。

D) 第一国以外の国の有無は「<H3>Also spoken in:</H3>」要素の有無により判断できる。あった場合は，その国名 C は<h4><A HREF="･･･">C</A></h4>の文字列を分割することにより取得できる。他の属性情報は第一国と同様に表形式になっている

(2) 語族別目次ページと語族ページの特徴

A) 語族別目次ページにある各々の語族名にリンク設定されている各語族ページの URL は「･･･/show_family.asp?subid=XXXXX」の形式になっていて，5 桁数字の語族ページ番号「XXXXX」だけが可変である。

B) 語族ページの言語系統樹の節点のテキストは<dt></dt>を用いて表現されていて， <dt>･･･</dt>要素を分割することにより取得できる(節点のタイプの識別については 3.1 を参照)。

Table 1　**言語系統樹の節点のテキストの形式と特徴**

節点のタイプ	節点のテキスト	節点のテキストの特徴
言語グループ	言語グループ名(下位階層に含まれている言語の数)[注1] 例 : Afro-Asiatic(375)	a．最後が必ず右丸括弧「)」である b．左丸括弧「(」が必ず1つ以上含まれる c．左丸括弧「(」と右丸括弧「)」の間は必ず数字
言語	言語名 [言語コード] (第一国の国名)[注1] 例 : Awjilah [auj](Libya)　注1:下線付き部分はいずれも可変要素を指す	丸括弧 ( )と鍵括弧 [ ]は不変要素で，必ずそれぞれ1組以上含まれる

C) 語族ページの言語系統樹の階層構造は<dl></dl>を用いて表現されている。言語系統樹の節点のレベル(level)[9]は「節点のレベル = 節点のテキストを表現している<dt>･･･</dt>要素を囲んでいる<dl></dl>の繰り返しの数 + 1」によって取得できる。

3.3　データの取得処理

言語属性情報と言語系統情報の取得処理の全体の流れをそれぞれ Fig. 2 と Fig. 3 に示す。流れ図の中の定義済処理については図の右にその処理概要を示している。両処理はともに MS-Excel(97-2003 ブック形式) VBA マクロより実装し, 取得情報は Excel ブックに出力している。

言語属性情報としては言語名，言語コード，第一国の国名，分類およびその言語が話されているすべての国における別名と方言名を取得対象にし，取得情報は Excel ブックの 1 シート(行が言語別, 国別, 列が項目別の情報の形式)に出力した。言語ページの文字コードは多言語対応の Unicode で, Excel(97-2003 ブック形式)も Unicode に対応しているため，取得した言語名等は文字化けすることなく Web ページの表示通りに保存できる。

言語系統情報の出力は，Excel ブックの 1 シートを 1 語族にしている。また，行と列の順番の数がそれぞれ節点の出現順とレベルを表わしている。

4　データのチェック・変換・利用

Excel に出力したデータの正当性は，特に言語系統樹の節点の間の親子関係についてのチェック処理, つまり, Table 1 に示しているように，言語グループの節点テキストからその下位階層に含まれている言語の数が読み取れるため，その数とその下位言語をカウントした数とを比較することで，エラーチェックを行った。

また，Excel に出力したデータについて，次のようなデータ変換処理を行った。

A) Yamamoto-Data が ASCII コードであるため，取得したデータを Unicode から ASCII に変換した。

B) 言語系統データを Excel 表形式から木構造の表現に適した XML 形式に変換した。各々の言語系統樹の節点のテキストは Excel に出力した節点のテキスト

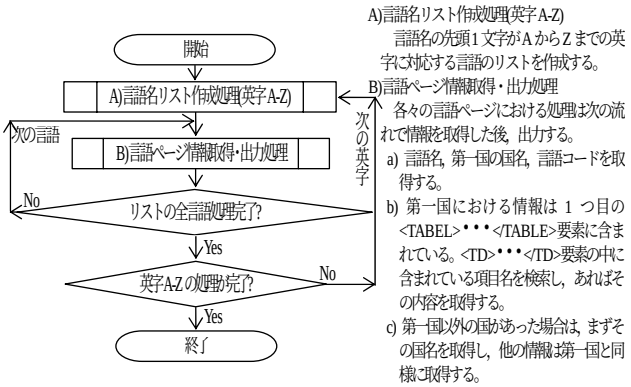

Fig. 2 **言語属性情報取得処理の流れ**

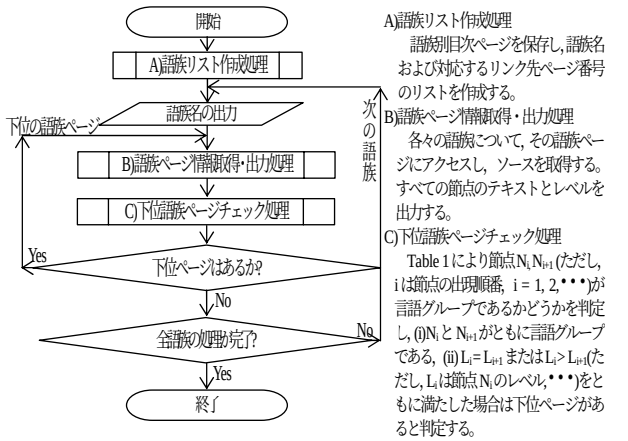

Fig. 3 **言語系統情報取得処理の流れ**

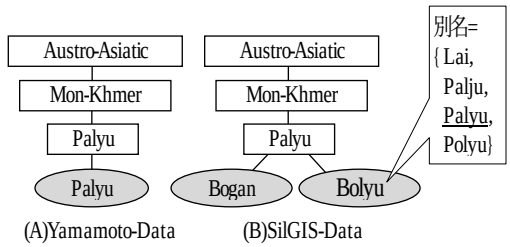

Fig. 4 **生成したデータによる言語同定の結果**

(Table 1 を参照)を文字列分解することにより得られた言語グループ名または言語名である。さらに，言語の節点に言語コード，国名を属性として付加した。

C) 言語属性データの言語の系統分類情報により，すべての言語の上記 B)の言語系統樹における経路を確定し，別名を言語の節点の属性として付加した。

生成できた属性情報付きの XML 言語系統データを言語同定手法[2]の処理に利用した。その処理では，言語系統樹における言語の経路および言語名 (言語の節点のテキスト)と別名(言語の節点の属性)により言語の同一性を判定する。結果として Fig. 4 にその 1 例を示す。(A)Yamamoto-Data の「Palyu」は(B)SilGIS-Data の「Bolyu」と同じ言語であることが確認された。

5 おわりに

本稿は *Ethnologue15th* Web サイトから言語の属性と系統の情報の取得およびデータ生成法を紹介し，その言語同定処理における利用についても述べた。インターネットは情報の宝庫ともいえ，そこから有用な情報を取得・変換し，必要なデータを作り出すことは今後益々その意義を呈すると思われる。本研究の成果である言語の属性と系統のデータは言語研究に有用なデータであることはいうまでもない。また，本研究で提案した手法は，インターネットからの情報の取得および活用の一手法として, 専門分野を問わず一般的に応用できると思われる。本研究で作成したソフトウェアは http://web.cc.yamaguchi-u.ac.jp/~k001wa/に公開している。

また，データ取得処理が，Web ページの構成が改変された場合にはそのままでは対応できないかもしれないなどの問題点があり，今後は特定のページ構成に限定しない，より汎用性のある手法を検討していきたい。

注と参考文献

[1] 呉靭，乾秀行，杉井学，松野浩嗣，「言語研究のための GIS データの生成について：*Ethnologue* GIS データを言語特徴の地図化に用いる一手法」，情報処理学会人文科学とコンピュータシンポジウム論文集，2007，pp.253-258

[2] Ren Wu, Hideyuki Inui, Manabu Sugii and Hiroshi Matsuno, “Language Identification for Generating GIS Data Used in Mapping Linguistic Features of the World's Languages”, Proceedings of ITC-CSCC2008, 2008, pp.153-156

[3] Gordon, R.G. (ed.), Ethnologue : Languages of the World, 15th edition, Dallas, SIL International, 2005
Ethnologue は国際 SIL(**I**nternational **S**ummer **I**nstitute of **L**inguistics)という言語研究団体が公開している出版物および Web サイト[4]のことを指す。*Ethnologue* は言語の開発と記録を促進するための言語学者の間における言語情報の共有を目的として創刊されたもので，言語に関する目録としては世界屈指の規模を有している。*Ethnologue15*th 書籍版の掲載情報を整理することにより言語の属性と系統のデータを作ることも可能である。

[4] http://www.ethnologue.com/web.asp

[5] Yamamoto-Data は弘前大学山本秀樹教授が公表されている「言語別語順データ」(山本秀樹，世界諸言語の地理的・系統的語順分布とその変遷，渓水社，2003)を指す。本研究では Yamamoto-Data に関連する言語系統データも必要であるが，その記述は割愛させていただく。

[6] SilGIS データは Shapefile(http://www.esrij.com/support/arcview3/material/shape/ shapefile.pdf を参照)の属性データ(dBase 形式)として提供されているため，Unicode 非対応で，フィールド長に制限がある。SilGIS データの文字化けはパッケージ製作時に Unicode のデータを直接 Unicode 非対応の dBase 形式に保存されたことによるものと思われる。

[7] 亀井孝ほか編著，言語学大辞典，三省堂，1996

[8] 世界諸言語の系統的分類は，実際は情報が少ないために語族としてまとめることができない言語が多い。*Ethnologue15th* の 108 語族の分類は一つの解釈である。

[9] 斎藤信男，西原清一，データ構造とアルゴリズム，コロナ社，1998

[10] 猿橋大，詳解 HTML タグ辞典，秀和システム ，2008

[11] https://www.microsoft.com/japan/msdn/web/server/asp/asptutorial.aspx

---

著者略歴

呉 靭（ご じん）
◎現在の所属：山口大学大学院理工学研究科(博士後期課程 D2)
◎専門分野：情報学，言語情報学

杉井 学（すぎい まなぶ）
◎現在の所属：山口大学メディア基盤センター
◎専門分野：情報学

乾 秀行（いぬい ひでゆき）
◎現在の所属：山口大学人文学部
◎専門分野：エチオピア少数言語の記述研究，言語類型論

松野 浩嗣（まつの ひろし）
◎現在の所属：山口大学大学院理工学研究科
◎専門分野：生物学，防災科学，言語学等他分野への情報科学の応用