GISを用いたデータベース構築に向けて*

乾　秀行

(山口大学)

inui@yamaguchi-u.ac.jp

0　はじめに

本稿は、言語研究に役立つ様々なデータをGIS（Geographical Information System)を用いてデジタル言語地図に関連づけるために、現在行っている作業について説明するものである。すでに乾(2006a, 2006b)において、GISを用いた言語研究の可能性、およびエチオピアでのフィールド調査した言語データ（語彙、文法、音声、映像の他、調査地点の緯度・経度情報等）のデータベース化に関する報告を行っている。そこで本稿ではその後どのような作業を行ってきたかについて報告する。

我々のプロジェクトでは、その後大きく分けて2つの作業を行ってきた。1つは「世界言語地図作成への取組」であり、もう一つは検索システムとして「XMLを用いた検索プログラムの研究開発」である。後者は本報告書所収の杉井氏および高橋氏の論考を参照していただくとして、ここでは前者について取り上げたい。

1　世界言語地図作成への取組

今までエチオピアの言語だけで行っていたGISによるデジタル地図化を世界言語全体に拡大するための取組は、筆者が平成18年度～21年度科学研究費補助金基盤研究(C)「地理情報システムによる世界諸言語の言語類型地理論的研究」代表山本秀樹（弘前大学）（課題番号18520288)に研究分担者として参加し、エチオピアの言語地図を作成した方法で世界言語地図作成に協力したことによる。ところが、これが思いの外時間を要する作業となった。エチオピアの言語地図

*本稿は、平成16～22年度科学研究費基盤研究(B)「オモ・クシ系少数言語の調査研究及び地理情報システムを用いたデータベース構築」代表乾秀行(山口大学)(課題番号16401008, 19401023)による研究成果の一部である。

作成の場合と同様、山口菱洋システム（http://www.yrs.co.jp）に依頼して、ベクトル形式で言語の範囲を1つ1つポリゴンという面で区切ってデジタル言語地図を作成するのであるが、紙媒体の言語地図を参考にしながらの作業は、元々の紙の地図の中に含まれる様々な情報の取捨選択が難しく、多くの部分で判断に迷う事例が出てきてしまった。不要な情報を削除し、とりあえずポリゴン化までの作業はほぼ終了し、そのポリゴンに言語名を関連づける作業に現在移行している。

それに関連して、我々の取組に先行して、すでに世界言語を対象にGIS向けのデジタル世界言語地図を作成し販売しているSIL（Summer Institute of Linguistics）の関連会社（GMI,Global Mapping International）のWLMS(The World Language Mapping System)を購入し、その評価を行った。これはSILが約4年に1回改訂している“Ethnologue”に準拠したものである。もしこの言語地図が我々の研究目的の実用に耐えうるならば、何も苦労をして新たに言語地図を作る必要はない。

まず手始めにSILの世界言語地図と我々がそれまで作成した世界言語地図をGISソフト（ArcGIS ver.9.2）上で重ね合わせてみたところ、世界のあらゆる地域で全くと言っていいほど言語の境界線が一致しなかった。ソースとした言語地図が異なっていることが主な原因と思われる。だたし、そのこと自体はそれほど大きな問題ではない。人は移動するものであり、国境線があって人の行き来がままならないならともかく、綺麗に境界線に沿って別の言語が話されていることなど現実にはありえない。大抵は何らかの行政区分に従い便宜的に境界線を作っているにすぎない。また世界のあらゆる地域で二言語併用、多言語併用地域が無数に存在するわけであり、境界線にそれほど重要な意味はない。したがって言語研究をする上でも地図上の分布がある程度解れば言語地図は目的を達成している面もあり、その意味ではたとえばHadpelmath et al.(2005)の言語特徴地図の表示方法でも質的・量的なデータさえ揃っていれば比較的単純な言語類型地理論的研究には十分耐えうるものとなるだろう。

ところがSILの世界言語地図の最大の問題は、たとえば一見すると北アメリカ大陸の言語地図がほとんど英語一色になってしまっている点にある。ただそれにはソフト上のちょっとしたからくりがあり、英語の部分を選択から外すとその下から隠れていた先住民族の言語が現れる仕組みになっている。しかし、その区分けを見ると、あろうことか、まるでアフリカ大陸の国境線を見るかのごとく、居住地が綺麗に四角く区切られているのである。つまり、現在の少数民族が暮らす保護地域であった。要するに、SILの言語地図はあくまで現在の言語の状態を表したものであり、コロンブスが訪れる前の先住民族の言語地図を全く考慮していないのである。少なくともこの地域の言語に関しては、何らか

の形でコロンブスが訪れる前の姿に復元しておかなければ、言語類型地理論的研究には利用できないことが判明した。ただし、著作権上デジタル言語地図をこちらの研究目的に合わせて勝手に改変することは許されていない。そこでやはり自らデジタル言語地図を作るという判断をした次第である。なお、その評価の詳細については呉他 (2007) を参照されたい。

SIL のデジタル世界言語地図の「属性データ」には今のところあまり言語特徴として利用できるものは入っていない。しかし我々が注目したのは、すべての言語を 1 対 1 で対応づけるための「ISO 639-3[1]」が入っている点である。この言語コードが一致しさえすれば、今まで言語研究をする上で言語名が一致しないためにデータの共有が思うように進まなかった点が解消されるからである。そこで我々のデジタル言語地図のポリゴンに言語名を関連づける際にこの「ISO 639-3」を使うことにした。現在その作業を進めているところである。

2 言語の ID 化の試み

2007 年 6 月にようやく「世界言語地図第 2 版（Asher & Moseley 2007）」が出版され、現時点で世界の言語を包括的に扱った言語地図としては最も信用できるものである。ここでは、この世界言語地図に載っている言語名に対して「ISO 639-3」による ID 化をする作業について、エチオピアのオモ・クシ系言語を例にとって説明する。

まず Ethnologue のホームページ (http://www.ethnologue.com/web.asp) に挙げられている言語名および言語コードを「系統分類（“Language Families”の項目）」にしたがって取り出す[2]。次に Asher&Moseley (2007) のエチオピアの言語地図 (pp.296-7) の横に載っている言語名リストをスキャナーで取り込み、言語名リストをデジタル化する[3]。

Ethnologue の分類では、エチオピアで話されているクシ系言語は全部で 24 言語、オモ系言語は 28 言語であった。一方、Asher & Moseley (2007) の分類によれば、別の番号[4]として割り振られている言語は、クシ系言語 21 言語、オモ系

[1]「言語コード」のことで、アルファベット 3 文字で表す。ちなみに日本語は「jpn」である。Ethnologue のホームページで確認できる。

[2]Ethnologue の言語リストの欠点は単純なアルファベット順に並んでいる点である。特に国別言語リストでは、系統関係を全く無視してアルファベット順に並んでいるので、照合作業をする上では極めて効率が悪い。系統分類を利用することでかなりその欠点が解消される。

[3]今回はエチオピアに限っているので、目視してエチオピアで話されている言語だけを選ぶ作業過程を入れた。世界言語を対象にする場合は、系統ごとに処理すればよいので、そのような作業過程は省かれる。

[4]方言差は番号の後にアルファベットが付いている。たとえばクシ系最大言語のオロモ語は 53 番であるが、53a〜53n までの 14 に下位区分され、地図上でも区分けされて表示されている。

言語34言語で、両データにはずれが生じている。

そこで両データの言語名を比較してみることにする。コンピュータ処理を行う場合、言語名の完全な一致があれば問題ないのであるが、言語名は統一されていないのが一般的で、複数の呼び名が存在することはよくある。今回の場合、言語名が完全に一致するデータは、クシ系12言語、オモ系14言語で、ちょうど5割の一致率であった。ただし、母音や子音表記による一部不一致は目視することで言語名の同定はそれほど難しい作業ではない[5]。そういった例を含めると、一致する言語はクシ系は8言語、オモ系は12言語それぞれ増えて、20言語、26言語となり、SILの分類の側から見て一致できない言語はクシ系で4言語、オモ系で2言語に減った。

ところでこの不一致の原因は、Ethnologueでは言語として数えているものをAsher & Moseley (2007)では、方言差と捉えてしまっていたり、またその逆である場合である。

まずクシ系言語から見ていく。Ethnologueに挙げられている「Bussa([dox])」、「Gawwada([gwd])」、「Tsamai([tsb])」の3言語が不一致なのは、Asher & Moseley (2007)では「Dullay」として表記されているためである。つまり、Ethnologueの系統樹を上に辿っていくとこれらの3言語の上のノードには「Dullay」がある。このようにEthnologueの側が細かく分類しているものに関して、系統樹の上の項目を見に行くようなプログラミングをすることで自動的に照合が可能であることがわかった。クシ系言語でもう一つ一致しない「Dirasha」は、その上のノードを見ると、「Konso-Gidole」となっている。次に“Alternate names”を見ると、「Gidole」の別名があることがわかる。しかしながらAsher & Moseley (2007)の地図にはどちらの名前も言語名として挙げられていないので、コンピュータ処理で自動的に照合することはできなかった。つまり、Asher & Moseley (2007)ではこの言語は「Konso」の一方言として処理されていると考えられる。逆にAsher & Moseley (2007)の方が詳しく分類をしている場合は、Ethnologueの“Dialects”の項目をチェックすることで自動的に照合できる。たとえばAsher & Moseley (2007)に挙げられていてEthnologueに見当たらない「Timbaro」という言語は、Ethnologueでは「Kambaata」の方言であると「Kambaata」の“Dialects”の欄に記載されている。

次にオモ系の言語についても見ておく。こちらの不一致は2言語である。まずEthnologueの「Shekkacho」の項目を見ると、“ Alternate name”に「Mocha」というのが挙がっているので、Asher & Moseley (2007)との照合が成功する。一

つまり別のポリゴンとなる。

[5]Ethnologueのデータには“Alternate names”という項目があり、別名リストが付いているので、目視したり記憶していなくても照合作業は自動化できる。

方、「Ganza」の方はAsher&Moseley (2007)の言語地図では、エチオピアではなくスーダンで話されていることになっていたので、最初からこの言語リストに挙げていなかった。通常のコンピュータ処理では不一致にならないであろう。

このような作業を世界言語に関して行うことで、おそらくAsher & Moseley (2007)の言語名のID化は実現可能であろう。ただし、今回はエチオピアのオモ系・クシ系言語に特化して説明をしたので、言語に関する予備知識が多少あるおかげで、言語の一致はそれほど難しい作業ではなかった。しかし、これが世界言語を対象にする場合、目視や知識に頼らなくてもいいような、自動化して処理する方法の精度を上げる必要がある。

最後にクシ系言語、オモ系言語の照合結果を表1、表2にそれぞれ示しておく。「−」は該当する言語が見つからない場合である。また、「A & M」の欄の「/」は、Asher & Moseley (2007)で別の言語として言語名が2つに分けられている場合である。「一致or不一致」欄の◎は言語名の完全一致、○は一部不一致、△は分類レベルの違いから一致可能、×は不一致をそれぞれ表す。なお、クシ系言語のうち「Somali」はエチオピア国内東部で話されている大言語であるけれども、Ethnologueの国別分類では主に話されているソマリアの方に分類されていた。前述のGanzaの例の逆のケースである。

表 1: Cushitic(Ethiopia)

言語名 (Ethnologue)	言語コード	言語名 (A & M)	一致 or 不一致
Xamtanga	[xan]	Khamtanga	○
Awngi	[awn]	Awngi	◎
Kunfal	[xuf]	Kunfal	◎
Qimant	[ahg]	kemant	○
Bussa	[dox]	—	△
Gawwada	[gwd]	—	△
Tsamai	[tsb]	—	△
Alaba	[alw]	Alaba	◎
Burji	[bji]	Burji	◎
Gedeo	[drs]	Gedeo	◎
Hadiyya	[hdy]	Hadiyya	◎
Kambaata	[ktb]	Kambaata	◎
Libido	[liq]	Libido	◎
Sidamo	[sid]	Sidamo	◎
Dirasha	[gdl]	—	×
Komso	[kxc]	Konso	○
Oromo, Borana-Arsi-Guji	[gax]	Oromo	○
Oromo, West Central	[gaz]	Oromo	○
Oromo, Eastern	[hae]	Oromo	○
Afar	[aar]	Afar	◎
Saho	[ssy]	Saho	◎
Arbore	[arv]	Arbore	◎
Baiso	[bsw]	Bayso	○
Daasanach	[dsh]	Dasenech	○
Somali	[som]	Somali	× (Somalia)
—	—	Timbaro	△
—	—	Dullay	△

表 2: Omotic(Ethiopia)

言語名 (Ethnologue)	言語コード	言語名 (A & M)	一致 or 不一致
Dizi	[mdx]	Dizi	◎
Nayi	[noz]	Nao	○
Sheko	[she]	Sheko	◎
Yemsa	[jnj]	Yemsa	◎
Chara	[cra]	Chara	◎
Bench	[bcq]	Bench	◎
Dorze	[doz]	Dorze	◎
Gamo-Gofa-Dawro	[gmo]	Gamo/Gofa	○
Melo	[mfx]	Malo	○
Oyda	[oyd]	Oyda	◎
Wolaytta	[wal]	wollaitta	○
Kachama-Ganjule	[kcx]	Kachama	○
Koorete	[kqy]	Koyra	○
Zayse-Zergulla	[zay]	Zayse-Zergulla	◎
Male	[mdy]	Male	◎
Basketo	[bst]	Basketo	◎
Anfillo	[myo]	Anfillo	◎
Boro	[bwo]	Bworo	○
Kafa	[kbr]	Kefa	○
Shekkacho	[moy]	—	△
Bambassi	[myf]	Mao of Bambeshi	○
Ganza	[gza]	Ganza	× (Sudan)
Hozo	[hoz]	Hozo-Sezo	○
Seze	[sze]	Hozo-Sezo	○
Aari	[aiz]	Aari	◎
Hamer-Banna	[amf]	Hamer/Banna	○
Dime	[dim]	Dime	◎
Karo	[kxh]	Karo	◎
—	—	Kullo	△
—	—	Konta	△
—	—	Doko	△
—	—	Gidicho	△
—	—	Mocha	△
—	—	Mao of Didessa	×

【参照文献】

Asher,R.E. & C.J.Moseley (2007) *Atlas of the world's Languages. 2 edition.* New York: Rougledge.

Haspelmath, M. & M.Dryer & D.Gil & B.Comrie (eds.) (2005) *The World Atlas of Language Structures. (Book with interactive CD-ROM)* Oxford: Oxford University Press.

乾秀行 (2006a)「地理情報システム (GIS) によるエチオピアのデジタル言語地図」*Cushitic-Omotic Studies 2006*, 山口大学, 1-7.

乾秀行 (2006b)「GIS を使ったクシ・オモ系言語研究」『一般言語学論叢』9, 47-58.

呉靚、山本秀樹、乾秀行、杉井学、松野浩嗣 (2007)「語順地図作成に必要なデータおよび語順地図に現れる語順分布」『一般言語学論叢』10, 31-49.