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—

呉 韌[†] 乾 秀行[††] 杉井 学[‡] 松野 浩嗣[†]
[†]山口大学大学院理工学研究科 [††]山口大学人文学部 [‡]山口大学メディア基盤センター

本研究は，世界諸言語の言語特徴の地図化に必要不可欠なGISデータの生成を目的とする．GISによる言語研究では研究対象言語の地理空間上の位置情報と言語特徴に関する属性情報が必要である．我々は，主に世界諸言語の位置情報を有するパッケージWLMS(The World Language Mapping System)のデータ(SilGISデータと呼ぶ)と山本が収集した語順研究のための言語属性データ(山本データと呼ぶ)を基に，言語研究に必要なGISデータの生成法を提案する．まずは，SilGISデータを利用する上での問題点について分析する．次に，SilGISデータを基に言語研究に利用できるGISデータの生成法を提案する．そして，山本の研究結果を基に生成されたデータの有用性について確認した上，SilGISデータの言語類型論研究への応用例を示す．

On Generating GIS Data for Language Studies
- A Method of Mapping of Language Characteristic from Ethnologue GIS Data -

Ren WU[†] Hideyuki INUI[††] Manabu SUGII[‡] Hiroshi MATSUNO[†]
[†]Graduate School of Science and Engineering, Yamaguchi University
[††]Faculty of Humanities, Yamaguchi University
[‡]Media and Information Technology Center, Yamaguchi University

This paper aims at generating GIS data that is necessary in mapping the language characteristic of all the languages of the world. In language studies using GIS data, the geographic information and the attribute information, which show the positions where the languages are spoken and various characteristics of the languages, respectively, are necessary and important. In this paper, we apply so-called SilGIS Data and Yamamoto Data to generate GIS data necessary for language studies, which are the data from package WLMS (The World Language Mapping System) including mainly the geographic information and the data collected by Hideki Yamamoto used in studying word order respectively. We firstly analyse SilGIS Data to point out the problems when used in language study, and then, based on SilGIS Data, we propose a method of generating GIS data necessary for language studies. After verifying the usefullness of the data generated by our method based on Yamamoto Data, we show an application of SilGIS Data to language typologie study.

1. はじめに

近年，自然科学の研究分野のみならず，人文社会系の研究分野においてもGIS（地理情報システム）のニーズが高まっており，言語研究への導入も始まっている[1,2]．

GISを言語研究に利用するにあたって，まずは研究対象言語の地理空間上の位置情報(言語の話されている地域の地理情報など)が必要である．言語特徴に関するさまざまな属性情報も欠かせない[2]．これらの空間データと属性データを解析・視覚化することによって，多元的な分析や判断が可能になる．我々は，GISの時空間検索機能を活用し，数理的処理手法も取り入れることによって，言語研究の中でも言語類型論的観点から世界諸言語を考察することを試みている．

言語類型論研究は様々な言語現象に基づき類型的特徴を考察し，言語普遍性を発見し，理論化していく学問で，従来は地理的分布や歴史的系統上の分類(語族)等の要因は捨象された状態で研究が行われてきた．しかし，近年多くの言語現象に地理的，系統的な要因が大きく関係していることが次第に明らかになりつつある[2]．このような地理的分布を重視する言語研究では言語地図が必要である．GISが言語地図の作成の強力なツールとなるであろうと期待され[1,2]つつも，まだ一般的には普及していないのが現状である．その原因の一つは言語地図の作成に利用できるGISデータの入手の困難さにある．筆者らの管見の及ぶ限り，現在一般に利用可能なそのようなGISデータはまだ少ない．

そこで本研究では，世界諸言語の言語特徴の地図化に不可欠なGISデータの生成手法を検討し，言語特徴の地図化の応用例を示す．言語特徴としてはまず世界諸言語の語順に焦点を絞ることにする．

語順に関する言語地図を作成するには，GIS属性データとして，言語を特定できる言語名や語順のタイプなどの項目が必要である．山本[3]は世界諸言語から幅広くデータを集め，2,932言語の語順データ(以下「山本データ」と称する)を収集した．本研究では山本データに基づいて，語順タイプの地図化データの生成を検討していく．

一方，言語に関するGIS空間データは，GIS向け

の世界諸言語の空間・属性データパッケージとして，WLMS (The World Language Mapping System) の最新版 (3.2.1) [4]が一般発売されている．WLMS は，その発売元と Ethnologue (書籍版[4]および Web 版[5]) を通して言語研究資源を提供している夏期言語研究所 (Summer Institute of Linguistics) との協力のもとでできた成果で，Ethnologue で掲載されている言語に関する情報および言語地図の作成に使ったものをデジタルデータとして提供していると思われる．WLMS (以下「SilGIS データ」と称する) には Ethnologue 第 15 版の 7,299 言語の言語名や言語が話されている国，また言語の別名や方言など多数の項目が含まれている．また，言語の空間データとしては面と点のベクトルデータが含まれている．

GIS 空間データとして既存のものが利用できれば，メリットは大きいが，筆者らの知る限りにおいて，世界諸言語の GIS データが一般発売されているのはこのほかに例がない．

しかし，SilGIS データを語順研究のために利用するには問題点がある．まず山本データと SilGIS 属性データとの言語の同定が必要である．言語名は言語学者によって同じ言語が異なる言語名で呼ばれていたり，異なる言語が同じ言語名になっていたりすることがしばしばある．同じ言語でも山本データの言語名と SilGIS 属性データのそれとは必ずしも一致しない．この点は，松本も語順データベースを構築した際の問題点として挙げている[7]．

また，SilGIS 空間データはそのデータ構築の目的および角度により，言語研究にそのままの形では利用できない情報が含まれている．我々は，SilGIS データが言語研究に直接利用できなくても，改編・加工すれば，言語研究に必要な GIS データを作り出すことが可能であると考えた．

そこで，本稿では，第 2 章において SilGIS データを利用する上での問題点を述べた上で，第 3 章において SilGIS データを基に言語研究に利用できる GIS データの生成法について述べる．それから第 4 章において山本の研究結果[3]を基に生成されたデータの有用性について確認し，言語類型論研究[8]への応用例を示す．最後の第 5 章において，今後の展望について述べる．

2. GIS データ生成の問題点

2.1 山本データと SilGIS データの形式

山本データ (表 1) には全部で 2,932 レコード (1 レコードが 1 言語) が含まれている．「言語名」順にソートされていて，「No」は筆者らが説明の便宜上つけたレコード通し番号である．言語の語順タイプとしては「節語順」や「接置詞」などが入っている．「国」はその言語が主に話されている国の国コードが最多で 3 つまで入っている．その国コードは山本独自のコード体系に従っている．

一方，SilGIS データには属性データと空間データがある．言語の空間データとして，ポリゴンフィーチャーと呼ばれる面の図形データとポイントフィーチャーと呼ばれる点の図形データの二種類が提供されている．ポリゴンフィーチャーデータの名称は「langa」(以下「langa データ」と称する) である．面および点の両データの構造は同じであるため，以下では langa データに限定して述べることにする．

langa データはシェープファイル[9]の形式で提供されている．シェープファイルはメイン・ファイル，インデックス・ファイル，属性ファイル (dBASE ファイル) という 3 つのファイルから構成される．langa データを例に，シェープファイルの構造を図 1 に示す．

図 1 に示すように，空間データの 1 レコードが一つの位置情報を示し，属性データおよび空間データのレコード間は 1 対 1 の関係をもつ．

SilGIS データには langa データのようなデータのほかに，ファイル名が eth_wlms.dbf となっている dBASE ファイル (以下「eth_wlms 属性データ」と称する) が単独で提供されている．それは言語の属性情報のみが格納されている表形式のもので，空間データと直接は 1 対 1 の関係をもっているものではないが，langa 属性データと同じ構造になっていて，フィールドが langa 属性データのものを包含している．eth_wlms 属性データの形式を表 2 に示す．表 2 に示すように，「ID_FIPS」や「言語名」を含めた 16 のフィールドが両者共通で，それに加えて言語の別名や方言，話者人口など多数の情報が eth_wlms 属性データに追加されている．

表 2 にある共通フィールドについては，「ID_FIPS」がレコードを一意的に識別フィールド (以下「キー」と称する) となっている．「ID_FIPS」の値は言語コード (言語の一意的識別子となる国際標準の最新コード体系 ISO639_3[5][10] に従っている．以下「3 文字言語コード」と称する) と国コード (FIPS 10-4 国コード[11] 体系に従っている．以下「2 文字国コード」と称する) をハイフン (-) で接続したもので，例として「jpn-JA」や「jpn-US」のような文字列となっている．また，説明の便宜上，筆者らが「言語名」および「ID_FIPS」の順にソートした場合のレコード通し番号を「No」とした．

表 1: 山本データ

No	言語名	節語順	接置詞	属格	形容詞	…	国	…
12	ABUN:MADIK	SOV				…	4II	…
151	ARANDA, WESTERN	SOV	PO	NG	NA	…	5AU	…
1038	JAPANESE	SOV	PO	GN	AN	…	4JA, 1US, 4SR, etc	…
1733	MONPA, CENTRAL	SOV	PO	GN	AN/NA	…	4ID	…
2099	POMO:SOUTHEAST	SOV	PO	GN	NA	…	1US	…

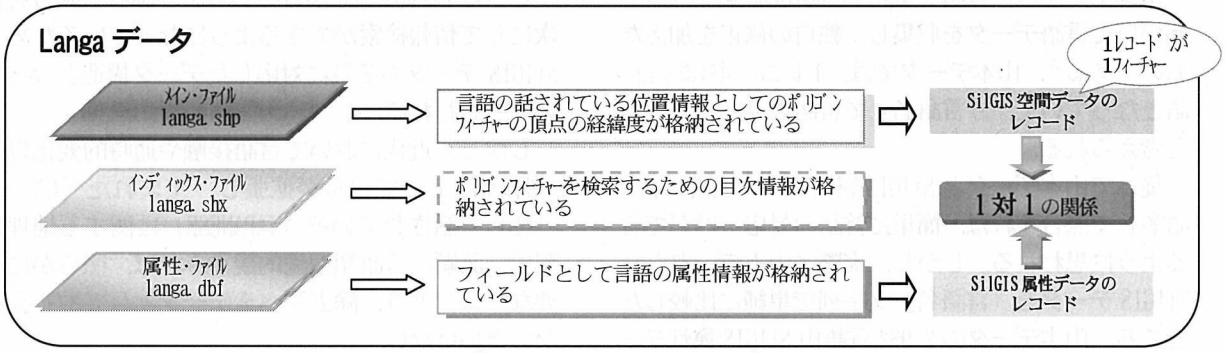

図 1:シェープファイルの構造:langa データの構成

表 2: eth_wlms 属性データ

No	…	ID_FIPS	…	言語名	…	…	別　名	…	方　言	…
30		kgr-ID		ABUN	…	…		…		…
532	…	are-AS	…	ARRARNTA, WESTERN	…	…	Aranda, Arunta	…	Western Aranda, Akerre (Akara), Southern Aranda	…
3901	…	jpn-CA	…	JAPANESE	…	…		…		…
3908	…	jpn-JA	…	JAPANESE	…	…		…		…
3920	…	jpn-US	…	JAPANESE	…	…		…		…
7558	…	pom-US	…	POMO, SOUTHEASTERN	…	…	Lower Lake Pomo	…		…
9389	…	tsj-CH	…	TSHANGLA	…	…	Sangla, …, Central Monpa	…		…

表 1 の No.12 に対応 (部分一致)

表 1 の No.151 に対応 (完全一致)

3 文字言語コード

2 文字国コード

表 1 の No.1733 に対応 (完全一致)

langa 属性データと eth_wlms 属性データの両方に含まれているフィールド

eth_wlms 属性データにのみ含まれているフィールド

eth_wlms 属性データはキーが langa データのような SilGIS 属性データと一致しているため，間接的に SilGIS 空間データのレコードとのリンクが可能である．以下では，「SilGIS 属性データ」をいう場合は表 2 の「eth_wlms 属性データ」にある共通データも含む．SilGIS 属性データの「言語名」は Ethnologue 第 15 版の第一言語名 (Primary langague name) [5] になっている．言語の別名 (Alternate names) [5] や方言 (Dialect names) [5] は eth_wlms 属性データの「別名」や「方言」に入っている．別名や方言は一つの言語につき複数存在する場合があるため，表 2 の「No」が 532 のレコードのように，複数の名称をカンマ(,)で接続した文字列となっている．

2.2 言語の対応付けにおける問題点

山本データは，言語を概ね Ethnologue 第 12 版に基づいて言語データを収集し，独自の修正を加えたものである [3]．山本データでは，1 レコードは 1 言語となっていて，「言語名」で言語を識別していると考えられる．

従って山本データと SilGIS 属性データとを「言語名」で照合すれば，簡単に言語の対応づけができるように思われる．しかし，実際に山本データと SilGIS データを「言語名」文字列で単純に比較したところ，山本データの 2,932 言語中 SilGIS 属性データから見つかったのが約 1,850 言語で，実に約 37% の言語がすぐには対応づけできない状況であった．

その一般的な理由として考えられるのは，「言語名」に入れるべき第一言語名が各々の言語学者の知見が入るため必ずしも一致しないことが考えられる．言語学者によっては，Ethnologue が別名や方言として挙げている名称を第一言語名としていることなどはしばしばある．

山本データは「言語名」としては Ethnologue 第 12 版の第一言語名に概ね準拠しているとされているが，Ethnologue の改訂によって，言語数が大幅に増えた (第 12 版では 6,528 言語，第 15 版では 7,299 言語) ことにより第一言語名に変更があったと思われる．

また，両データにおいて，同じ言語が異なる言語名になっている場合と異なる言語が同じ言語名になっている場合を調査したところ，それぞれ言語名が重複している言語があることを確認している．

このように，「言語名」文字列の重複がある以上，このままでは山本データと SilGIS 属性データの言語の対応づけはできない．「言語名」のほかに，言語を一意的に識別できるものが必要である．山本データではこのようなフィールドが存在していないため，付加する必要がある．

2.3 データ構造上の相違点

表 2 からわかるように，SilGIS データの 1 レコードには一つの言語が一つの国で話されている地理情報とリンクする属性情報が格納されている．つまり，一つの言語につき，SilGIS データに複数のレコードが存在する場合がある．一つの国についても同様である．例として，表 2 の 3 文字言語コードが「jpn」の「JAPANESE」言語は，2 文字国コードが「CA」や「JP」，「US」などを含めた 25 カ国で話されている地理情報とリンクしている可能性がある (前述のように，表 2 のレコードは必ずしも空間データに対応しているわけではない)．

これは Ethnologue が言語名および国名の両方を目次にして情報検索ができるようになっているため，SilGIS データもそれに対応したデータ構造となっていると思われる．

しかし，近代において言語接触や通時的変化以外の原因によって言語が拡張し，生まれた「CA」や「US」で話されている「JAPANESE」に関する地理情報は，語順の言語類型論的観点みれば，明らかに不要なものであり，除去すべきデータとして取り扱わないといけない．

3. 山本データに基づく GIS データの生成

3.1 言語コードの付加

Ethnologue では言語名の重複問題を避けるため，言語の一意的識別子となる 3 文字からなる言語コードが以前より導入されており，第 15 版は国際標準の最新規格である「ISO639_3」というコード体系に準拠している [5]．前述のように，SilGIS 属性データの「ID_FIPS」(表 2) のハイフン (-) の前にある 3 文字がその言語コードとなっている．

山本データの言語に対し，その言語名を eth_wlms 属性データから検索し，同じ言語と同定できる言語の 3 文字言語コードを山本データに付加することを考える．検索するフィールドとしては「言語名」のほか，「別名」や「方言」を対象とする．

3.2 必要な GIS データの選別

SilGIS データには言語に関する位置情報としては不必要なものが含まれている．例として，表 2 では「JAPANESE」言語の 25 レコードのうち，必要なレコードは「ipn-JA」だけで，その他のレコードは語順研究では不要である．不要なものを除去するよりも必要なものを選別する方が容易であるので，本稿上

で述べたような一つの言語に対応する複数の地理情報から必要なものだけを選ぶことにする.

Ethnologue 第 15 版ではすべての言語に対し，第一国 (Primary Country)[5]が示されていて，それは普通その言語の発祥地または最も多くの話者がいる国とされている．言語に対するその第一国の地理情報が言語研究に必要な地理情報と考えられる.

一方，山本データでは一つの言語に対して「国」に主に話されている国が最多 3 つまで入っている．山本[3]は「本研究では言語に関して地図や地理的分布という場合は，言語学における言語地図の作成で通常行われているように，可能な限り原住の言語を対象にしている．たとえば，オーストラリア，ニューギニア，アフリカ大陸，アメリカ大陸などにはヨーロッパの言語をまったく入れない形で，それぞれ地元の在来の言語を対象にしている」と記している．この点は，筆者らの認識と合致しているといえる．山本データの「国」の 3 つの国のどれが第一国なのかについて明記はされていないが，筆者らは前記山本の知見により「国」の 1 つ目の国が第一国に当たると推定するのが妥当であると考える.

また，山本データは Ethnologue 第 12 版に概ね準拠したものに山本独自の修正を加えている．本研究ではこの山本データに従って言語に関する必要な位置情報の選定処理を行うことにする.

3.3 データ処理手順

(1) eth_wlms 言語一覧の作成

表 2 の「ID_FIPS」に含まれている 3 文字言語コードと「言語名」をフィールドとする言語一覧表を作成する(以下「eth_wlms 言語一覧」と称する).

(2) 言語名重複言語の分離

前述のように，山本データおよび SilGIS 属性データの両方において異なる言語の言語名が同じ場合がある．このような言語は「言語名」だけでは特定できないため，言語名重複のない言語と分離して処理する必要がある.

山本データを「山本データ 1」と「山本データ 2」に分離する．「山本データ 2」には次のような言語が含まれる.

・山本データの言語名が重複している言語

・山本データでは言語名が重複していな eth_wlms 言語一覧においてその言語の言語名が重複している言語

つまり「山本データ 1」には「山本データ 2」の言語を除く言語が含まれる．以下では，まずその「山本データ 1」を山本データとして処理する.

(3) 山本データと eth_wlms 属性データの言語名に使う文字や符号の統一およびデータ処理ツールで使用できない符号の除去

言語名称の文字列の要素として使われている文字や符号を統一する.

また，両データに「言語名」や「別名」，「方言」の文字列にダッシュ(‘)やダブルコーテーション(“”)が含まれていることがある．使用するデータ処理ツールによってはシステムエラーを起こす可能性があるため，予め削除しておく.

(4) 3 文字言語コードの特定

山本データの言語名に対し，表 2 の eth_wlms 属性データの「言語名」，「別名」および「方言」の文字列を対象に，文字列検索処理を行う．処理後のデータを次の 3 つのレベルに出力する.

・レベル 1:完全一致データ
言語の同定ができ，3 文字言語コードを山本データに付加できた言語

・レベル 2:部分一致データ
言語名の文字列の部分的一致が認められたが，言語の特定は更なる検討が必要な言語

・レベル 3:不一致データ
完全一致および部分一致以外の言語

以下ではレベル 1 のデータを対象とする.

(5) 山本の国コードの変換

山本データの「国」に含まれている 1 つ目の国コードを 2 文字国コードに変換する.

山本データでは「国」が空白となっている言語もあるが，そのよう言語はここで処理対象外とする.

(6) データ形式の統一

山本データに付加された 3 文字言語コードと変換後の 2 文字国コードを基に，SilGIS 属性データの「ID_FIPS」と同じ形式のフィールドを山本データに追加する.

(7) 必要な SilGIS 空間データの抽出

山本データに対し，「ID_FIPS」をキーに langa 属性データと結合することにより，対応する langa 空間データを抽出する.

4. 処理結果および地理的分布図の作成

前述の手法に従って Microsoft Office Access2007 を用いて処理して得られたレベル 1，レベル 2，レベル 3 のデータにはそれぞれ約 2,100 言語，220 言語，450 言語が入った．レベル 1 データとレベル 2 データの例を表 2 の吹き出しに示している.

さらに，レベル 1 のデータおよびポリゴンフィーチャー空間データ (langa 空間データ) によって GIS

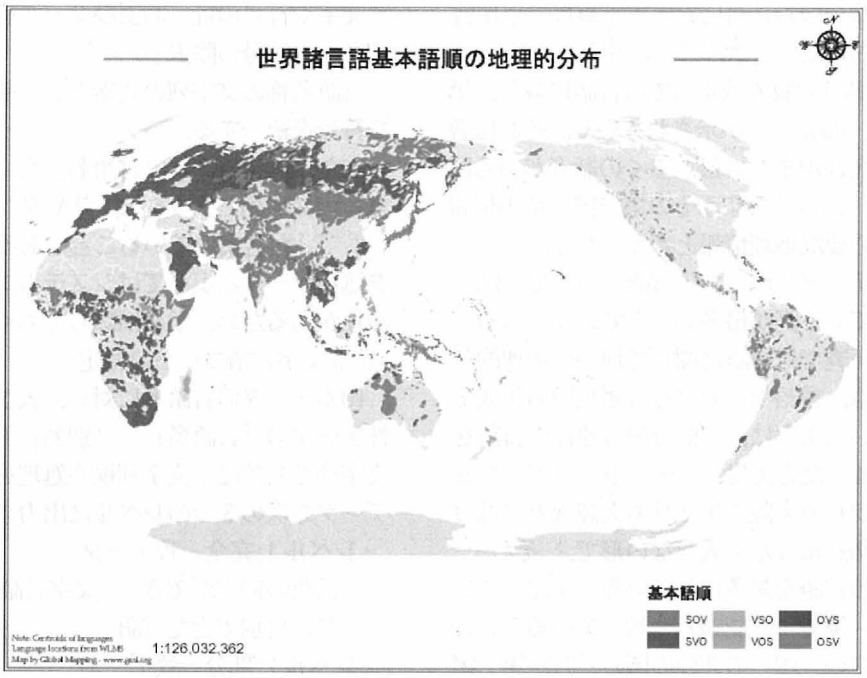

図 2:世界諸言語基本語順の地理的分布図

データを生成したところ，約 1,800 言語に関して言語地図の作成に使えるデータがえられた． GIS ソフト (ArcView9.1) を使って作成した基本語順 (山本データの「節語順」) の地理的分布図を図 2 に示す．

山本が指摘[3]したように，図 2 からは， SVO はヨーロッパ，アフリカ大陸，東南アジアの 3 つの地域に特にかたまって表れていて，それ以外の地域では，アジア北部，インド亜大陸等，広い範囲にわたって SOV 地域が連続していることが顕著に表れていることがわかる．

5. 終わりに

本稿では言語研究に利用できる GIS データを生成する手法を提案した．また，生成した GIS データを用いて言語特徴の一つである基本語順の地理的分布図を作成し，応用例として示した．

ここに提案した手法は山本データのような語順データだけでなく，ほかの言語特徴データにも応用できる．さらに，生成できた GIS データは従来の言語学者の知見を確認するような言語地図の作成に役に立つだけにはとどまらず，従来の研究方法では見出すことが困難な言語学分野での新しい発見につながる可能性がある．

なお，山本データと SilGIS 属性データとの言語の対応づけにおいてデータ処理の結果として部分一致および不一致の言語をどう同定していくかが，今後の課題のひとつである．

謝辞

本研究の遂行にあたり，貴重なご助言および山本データの電子版をご提供いただいた弘前大学山本秀樹教授，並びにデータ調査など多大なご尽力をいただいた山口大学理学部自然情報科学科の伊藤佳奈さんに，深い感謝の意を表する．

参考文献

[1] 池田潤: GIS と言語研究，一般言語学論叢，第 9 号，pp. 1-10，2006.
[2] 山本秀樹: GIS と言語類型論:世界言語地図に基づく言語研究，一般言語学論叢，第 9 号，pp. 31-40，2006.
[3] 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，溪水社，2003.
[4] http://www.gmi.org/wlms/
[5] Raymond G. Gordon, Jr.: Ethnologue: Languages of the World, 15th edition. SIL International, 2005.
[6] http://www.ethnologue.com/web.asp
[7] 松本克己:語順データベース，日本語学，13-1，1994.
[8] 松本克己:世界言語への視座-歴史言語学と言語類型論，三省堂，2006.
[9] http://www.esrij.com/support/arcview3/material/shape/shapefile.pdf
[10] http://www.sil.org/iso639-3/default.asp
[11] http://www.itl.nist.gov/fipspubs/fip10-4.htm