O-001

インターネット上の言語分布に関する調査

A survey on language distribution on the internet

新井 裕樹 † 中平 勝子 † 三上 喜貴 †
Yuki Arai Katsuko T. Nakahira Yoshiki Mikami

1 はじめに

情報社会と呼ばれる現在，人はインターネットを介して世界中のあらゆる情報へ瞬時にアクセスできる．しかし，その恩恵を享受することの可能な人と不可能な人との間に生じる情報格差（デジタル・デバイド）が問題視されている．木村忠正は『情報ネットワークへのアクセスを「もつ」「もたない」が，社会階層（国家）により大きく異なり，しかも，そうした階層間（国家間）の経済的格差，社会的格差が拡大する傾向にある』と述べている [1].

このような国別の情報格差の実態を測定評価するために，ITU の Digital Access Index[2] をはじめとして，ネットワーク環境，パソコン台数等の一次データから構成される様々な情報格差指標が開発されてきた．しかし，既存の指標の多くは，通信基盤の投資度合や充実度のみを考慮したものであり，インターネット上の言語の利用実態に関する格差を考慮に入れたものはない．情報格差が情報ネットワークへのアクセス機会の格差に起因することを考えれば，通信設備やパソコンの台数などの基盤を主体に格差を計測し，評価することは妥当とも言える．

しかし，その一方でそうした情報基盤の上を流通する情報そのものに対する関心も高まっている．世界情報社会サミット（World Summit on the Information Society: WSIS）のジュネーブで行われた会合では，情報社会に向けた共通ビジョンを定義した「基本宣言」が宣言された．そこで情報の公平な分配と全ての人による利用を保障し，また，多言語が使用されることを考慮しなければならないとインターネット上を流通する情報への言及が見られる [3].

実際の問題としてコンピュータ上で使用できる言語は限られており，文字コード標準が未確立の場合，現地ベンダー等は独自の非標準文字コードとこれに対応した専用フォントを開発することになり，南アジアや東南アジアの文字については特にこの傾向が見られる．文字コード標準が技術的に制定されているか否かで，母語による情報発信の可否が決定され，文字コード標準の有無によって格差が生じている [4]．また，文字コード標準が制定されていても，非標準コードをもとにしたフォントが使用されているため，情報の流通が阻害されている言語も存在する．

そこで，本稿ではインターネット上の多言語使用の実態を把握するためにインターネットにおけるページの使用言語を調査し，国別の公用語使用率の利用実態を明らかにする．調査は，e-Society プロジェクトで収集した Web ページのデータを元に，ページの TLD とサーバの所在地から国情報を判別し，独自開発の言語判定エンジンによって使用言語を特定し国別の使用ページを判定した．

† 長岡技術科学大学 Nagaoka University of Technology

2 先行研究

三上らの言語天文台プロジェクトは，言語間デジタルデバイドの解消を目指して，インターネット上の言語活動を観測する言語天文台を構築し，言語間デジタルデバイドの計測を行っている．インターネット上の言語活動を把握することで母語や，使用文字コードの現状について把握し，少数言語の参加を促す手段を見つけ出す活動を行っている [5].

星野らは，アジア・アフリカドメインにおけるネットワーク環境に関する調査研究の一環として，ネットワークの物理的基盤である通信パフォーマンスとサーバの設置状況を調査し，特にアフリカでは，国内設置の割合が非常に低いことを明らかにした．国外設置を必要とされる理由として，通信回線の設備の遅れ，電源の安定性に関する問題，サーバ管理に関する技術者の不足などの諸要因が指摘している [6].

童らは 107 億 Web ページに対し，言語分布，TLD 分布等の解析を行うと共に，日本語 Web ページを 1 ページ以上含む Web サーバから発信される約 3 億の Web ページに対して，Web サーバの地理的な位置を特定し地理上での分布・リンク特徴抽出を行っている [7].

以上のことから，これまでの研究では，TLD ごとの言語分布の調査やサーバの所在地の調査を行っているが，サーバの所在地ごとの言語分布の調査を行っていない．そのため，サーバの所在地別の言語分布，ccTLD 別かつサーバの所在地別の言語分布の調査を行う．

3 調査方法

3.1 調査対象となる Web ページ

調査対象となる Web ページは，e-Society プロジェクトによって収集した Web ページの URI 情報を元に収集した Web ページである．e-Society プロジェクトの Web ページは，2007 年 8 月までの 100 億超の収集により日本語を 1 ページ以上含むホストのみを収集したものである．収集した Web ページは，Basis Technology の言語判定システムで英語，日本語，中国語，フランス語，韓国語，スペイン語，ドイツ語，イタリア語，ロシア語，ポルトガル語，アラビア語の 12 言語として判別できなかった Web ページを対象とした．これは非標準文字コード使用ページの言語分布の調査のために，話者数が多く文字コードが標準で存在するページデータを予め除外した方が言語分布の把握することが容易であると考えたからだ．サーバ別の主体

のため，調査は，収集した 735,347 ページを対象として解析を行う．

3.2 解析システム

本稿の言語情報の計測は，情報格差観測のための統合管理システムの一部の機能を使用した．情報格差観測のための統合管理システムの概要を図 1 に示す．大きく入力部・出力部，およびそのインタフェースから構成される．また，継続的運用を意識したため，特別に管理部を設けている．本稿では，入力部の抽出フェーズのツールである LIM/G2LI と GeoIP を使用した．

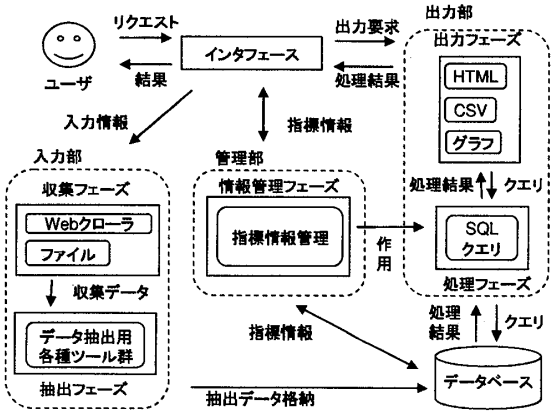

図 1 システム概要図

LIM（Language Identification Module）は，ウェブページやテキストデータの使用言語，使用文字体系，使用エンコーディング方式を自動的に判定できる言語判定エンジンである [8]．LIM は，N-gram を利用して言語判定を行っており，現在約 334 言語を識別できる．G2LI は LIM に Unique アルゴリズム，Frequency アルゴリズムの二つのアルゴリズムを追加し，Web 文書のための HTML フィルタリング機能を追加したものである [9]．

GeoIP は，GeoIP は，MaxMind 社が配布している IP アドレスからサーバの設置国コード，サーバの経度・緯度を求めるためのツールである [10]．

3.3 調査項目

言語分布の調査のために，以下の 4 項目の調査を行った．

・Web ページ言語分布

調査対象の Web ページの使用言語ごとのページ数について調査を行った．調査対象の言語分布の全体像を把握するために調査した．データ量が多いため，使用言語ページ数の多い上位の言語と出現した非標準文字コード使用ページの言語のデータのみ示す．

・ccTLD/gTLD 別の言語分布

調査対象の Web ページの ccTLD/gTLD 別の使用言語のページ数についてそれぞれ調査を行った．データ量が多いため，ページ数の多い上位 20 の ccTLD/gTLD 別の言語と出現した非標準文字コード使用ページの言語のデータのみ示す．国ごとにデータを分別するため，ccTLD 別の言語分布では.com などのジェネリックドメインの Web ページは除いている．gTLD 別の言語分布は ccTLD 別との相違を見るために調査した．

・サーバの所在地別の言語分布

調査対象の Web ページのサーバ所在地別の使用言語ごとのページ数について調査を行った．データ量が多いため，ページ数の多い上位 20 のサーバ所在地別の言語と出現した非標準文字コード使用ページの言語のデータのみ示す．

・ccTLD 別かつサーバの所在地別の言語分布

調査対象の Web ページの ccTLD 別サーバ所在地別の使用言語ごとのページ数について調査を行った．当該ドメインの元に置かれたサーバの所在国を調査し，ccTLD 別に見た自国内・国外サーバに存在するページ数を求め，そのページで使用されている言語が公用語か非公用語であるかを判断し，国内外サーバに置ける公用語のページ数を求めた．データ量が多いため，使用言語ページ数の多い上位 5 のデータと出現した非標準文字コード使用ページの言語のデータと特徴のあるデータを示す．国ごとにデータを分別するため，.com などのジェネリックドメインの Web ページは除いている．

4 調査結果

735,347 ページを対象として収集・調査した結果を以下に示す．

4.1 Web ページ言語分布

調査対象データの Web ページ言語分布を調査した結果を図 2 に示す．

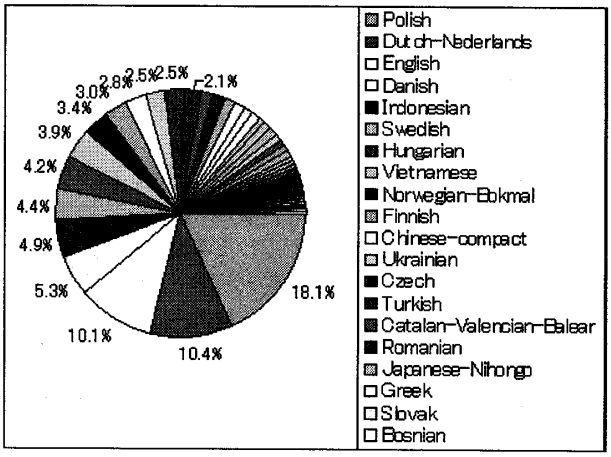

図 2 言語分布

図 2 から Web ページで使用言語が多いのは，ポーランド語，オランダ語などのヨーロッパ地域の言語やインドネシア語，ベトナム語などのラテン文字を使用したページが多いことが分かった．非標準文字コード使用ページの言語は，アルメニア語が 424 ページ，タミル語が 17 ページ，シンハラ語が 13 ページ存在したが，いずれも少数のページしか存在しなかった．

4.2 ccTLD/gTLD 別の言語分布

調査対象データの ccTLD 別の言語分布を調査した結果を図 3 に示す．

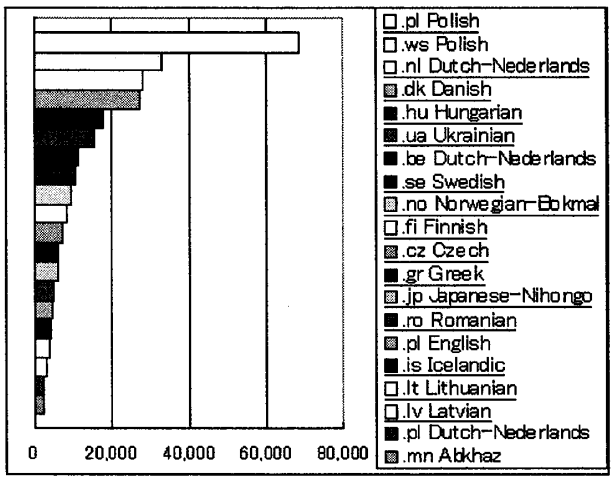

図 3 ccTLD 別の言語分布

図 3 の 15 の ccTLD 別の使用言語（下線のある言語）は，ccTLD で使用されている公用語が利用されていることがわかる．他の 5 個の ccTLD と言語の組み合わせについては，.ws ドメインでポーランド語，.mn ドメインでアブハズ語が多く使用されていることが分かった．非標準文字コード使用ページのの言語については，アルメニア語が.am ドメインにて 357 ページ，シンハラ語が.lk ドメインにて 13 ページと公用語に対するドメインで使用され，タミル語が.in ドメインと.lk ドメインそれぞれに 2 ページずつと.net などの gTLD にページが存在することが分かった．

調査対象データの gTLD 別の言語分布を調査した結果を図 4 に示す．

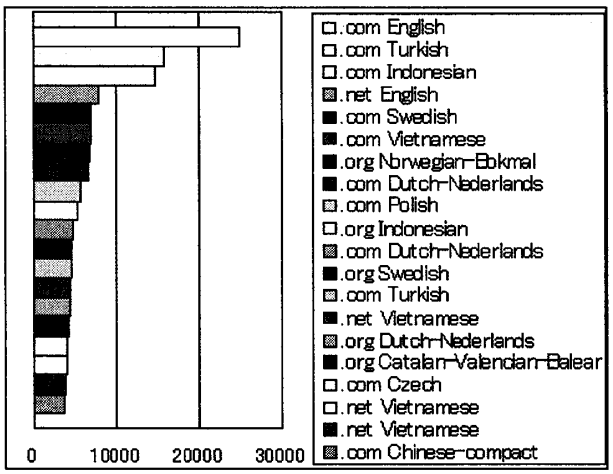

図 4 gTLD 別の言語分布

gTLD 別の使用言語は，.com ドメインにおける英語が 24,977 ページと一番多い結果となったが，ccTLD 別の言語分布に比べ，トルコ語やベトナム語が上位に多数にランクされる結果となった．これは，これらの言語で情報発信する場合，ccTLD よりも gTLD の方が好まれ利用されていることが示すものであると考えられる．非標準文字コード使用ページの言語については，アルメニア語が.net:50，.org:1，.gov:1，タミル語が.net:11，.org:1 ページ存在し，シンハラ語は gTLD 上には存在しなかった．

4.3 サーバの所在地別の言語分布

調査対象データのサーバの所在地別の言語分布を調査した結果を図 5 に示す．

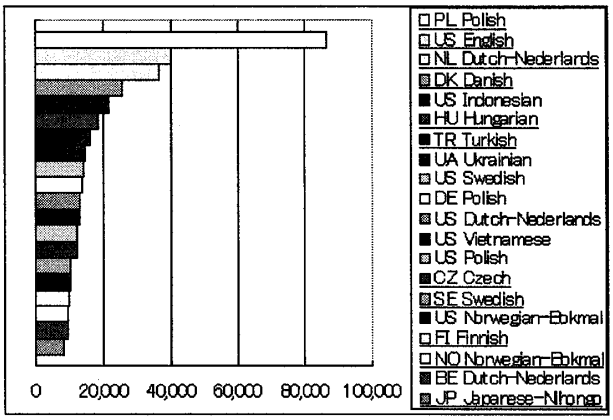

図 5 サーバの所在地別の言語分布

11 のサーバの所在地別の言語（下線がある言語）は，サーバの所在地で使用されると考えられる公用語が利用されていることが分かる．他は，主にアメリカで設置されたサーバ上で，インドネシア語，ベトナム語，ポーランド語，ノルウェー語などが使用されたページが多いことが分かった．非標準文字コード使用ページの言語は，アルメニア語がアルメニアサーバに 357 ページ，タミル語がインドサーバとスリランカサーバにそれぞれ 2 ページずつ，シンハラ語がスリランカサーバに 13 ページ存在した．

ccTLD 別の言語分布からはページの使用言語に対応した ccTLD が多いことが分かったが，それに対してサーバの所在地別の言語分布からはサーバの所在地としてアメリカが多い結果となった．これから推測すると，公用語による情報の公開用の ccTLD は情報発信者の母国 ccTLD を使用するが，実際にサーバが設置されている国はアメリカが多いと推測できる．

4.4 ccTLD 別かつサーバの所在地別の言語分布

表 1 サーバ所在地別の公用語ページ合計数

	国内	国外
公用語	209,741	20,216
非公用語	52,805	88,898

調査対象データの ccTLD 別かつサーバの所在地別の言語分布を調査した．表 1 は，各 ccTLD ごとにサーバの設置国を

表 2　ccTLD 別かつサーバの所在地別の言語分布

.pl	国内	国外	.ws	国内	国外	.nl	国内	国外	.dk	国内	国外
公用語	50,459	0	公用語	0	424	公用語	25,873	2,252	公用語	24,898	2,588
非公用語	7600	32,905	非公用語	0	34,736	非公用語	3,494	1,328	非公用語	2,233	172
.ua	国内	国外	.am	国内	国外	.in	国内	国外	.lk	国内	国外
公用語	14,069	1,424	公用語	357	0	公用語	35	0	公用語	15	0
非公用語	3,188	2,298	非公用語	66	10	非公用語	30	76	非公用語	1	11
.gr	国内	国外	.cc	国内	国外	.ms	国内	国外	.tv	国内	国外
公用語	3,544	4,070	公用語	0	0	公用語	0	0	公用語	0	238
非公用語	1,150	267	非公用語	0	124	非公用語	0	2,778	非公用語	0	305

求め，公用語で使用するページ数を調査したものをすべての ccTLD の値を合計したものである．

表 1 から調査対象の Web ページは国内に設置されたサーバ上で公用語による発信する情報量が多く，国内サーバ上では海外へ向けた非公用語による情報発信量は約 4 分 1 であることが判明した．

次に，各 ccTLD ごとの詳細なデータを示す．ここでは 12 のドメインの調査結果を選別し表 2 に示す．ページ数の多い上位 5 位ドメイン（.pl，.ws，.nl，.dk，.ua）と非標準文字コード使用ページの言語が存在する 3 つのドメイン（.am，.in，.lk），残りはある特定の特徴を持つ 4 つドメイン（.gr，.cc，.ms，.tv）を取り上げた．

ページ数の多い上位 5 位ドメインのうち.ws を除く 4 つのドメインでは，国内で公用語で書かれたページが一番多く，調査対象となったドメイン 109 個のうち 27 個がこのように国内サーバ上で公用語によって情報発信するページ数が多い国である．.ws ドメインは国外サーバで非公用語によって情報を発信しており，その約 95% がポーランド語で表記されているものであることが分かった．

非標準文字コード使用ページの言語が存在するドメイン（.ar, .in, .lk）は，どのドメインも国内サーバ上のみで公用語で情報を発信することが分かった（.in の 35 ページのうち 33 ページが UTF-8 で書かれたページで，専用フォントを使用したページは 2 ページ）．.gr のドメインは，他 ccTLD と比べ海外に設置されたサーバ上で公用語で情報発信するページが多いことが分かった．

.cc，.tv，.ms などの島嶼国では，自国にサーバを置かず非公用語による情報の発信が多いことが判明した．これからは，島嶼国では自国のドメインを海外に売ることで収益を得ていることを推測できる．特にツバルの.tv ドメインは"TeleVision"の略"TV"という二文字は放送業界でよく使われている．実際，.tv ドメインを目視したところ，放送局関連のホームページであることが確認できた．このようなことから，自国にサーバを設置せず非公用語による情報発信が多いことが推測できる．また，モントセラトの.ms ドメインではタイサーバ上でタイ語と英語によって主に情報が発信されていた．Web ページを実際に確認したところ"motor sports"に関するページであり，その略である"ms"をドメインとして使用したと推測でする．

5　まとめ

本稿ではインターネット上の多言語使用の実態を把握するためにページの使用言語を調査し，国別の公用語使用率の利用実態を明らかにした．インターネット上で使用されている言語は，ヨーロッパ諸国等のラテン文字を使用した言語が多く，非標準文字コード使用ページの言語の利用は限定的であった．非標準文字コード使用ページの言語のページの取得数が少量である理由は定かで無いが，この研究で収集した Web ページが少なかったことと，クロウリングの始点となる URL が日本語に偏っていたことであると推測する．今後はこの原因を追究・改善し，調査を続けたいと考える．

参考文献

[1] 木村忠正, デジタルデバイドとは何か，岩波書店（2001），p.10.

[2] International Telecommunication Union, http://www.itu.int/net/home/index.aspx

[3] World Summit on the Information Society, Declation of Principles, http://www.itu.int/wsis/docs/geneva/official/dop.html.

[4] 三上喜貴，文字符号の歴史 アジア編，共立出版（2002），p.205.

[5] 児玉茂昭，チュー ユー チョーン，三上喜貴，自動言語判定手法の開発とそれを利用したインターネット上の言語分布に関する調査，日本言語学会第 135 回大会，松本，2007.

[6] Katsuko T. Nakahira et. al, Geographic Location of Web Servers under African Domains, The 15th International World Wide Web Conference, Edinburgh, 2006.

[7] 童芳，平手勇宇，山名早人，全世界の Web サイトの TLD・言語分布・地理的設置位置の特定，日本データベース学会論文誌，Vol.7, No.1, pp.31-36（2008）.

[8] 中鉢欣秀，Gondri Nagy Janos 他，言語天文台を設立するための言語判定フレームワークの開発，第 171 回自然言語処理研究会，2006.

[9] Chew Yew Choong, 中平勝子, 三上喜貴, Language identification on text corpus using N-gram statistical classifiers, 平成 19 年電子情報通信学会信越支部大会, p.64.

[10] MaxMind co. ltd., GeoIP, http://www.maxmind.com/.