アジア及びアフリカの地域別ドメインにおける言語の分布状況について

児玉茂昭† 三上喜貴† Chew Yew Choong†
長岡技術科学大学†

1. 本研究の目的

発表者らが進める言語天文台プロジェクトは、2006年から年1回、アジア・アフリカ地域の地域別ドメイン(ccTLD)に存在するWebページにおける言語の使用状況を調査してきた。本発表では、この調査をもとに、アジアとアフリカの間で、英語などの大言語の使用状況、タイ語などの地域言語の使用状況にみられる差異を明らかにし、その差異の原因を考察する。

インターネットが重要な情報伝達手段になるにつれて、その上での情報格差、すなわちデジタル・ディバイドが、重要な問題となっている。ITU (International Telecommunication Union) のICT Development Index[1]では、様々な経済指標や情報基盤に関する調査が行われている。

しかし、情報の媒体である言語についての調査は少ない。特に、話者数の比較的に少ない言語の調査については、少ないインターネット上の情報を母語で利用することできないという言語的デジタル・ディバイドの実態の調査もデジタル・ディバイドの問題を考える上で重要であり、このことが、本調査研究を進める主要な動機となっている。

2. 調査方法

言語天文台が2009年に収集したWebページを対象として分析を行った。分析の対象とする地域ドメインは、アジア地域が、日本、韓国、中国などを除いた42ドメイン、アフリカ地域が、57ドメインである。収集の後、重複ページなどを除いて分析の対象としたページの総数は、アジア地域が7,079,867ページ、アフリカ地域が18,361,500ページである。

収集にはUbicrawler[2]を用い[1]、分析には、Chewが開発した言語判定エンジンであるG2LIを用いた。収集を行う際にUbicrawlerに与えた開始URIは、アジア地域が2,357個、アフリカ地域が1,906個である。また、G2LIは、315言語の判定を行うことができる。

3. 調査結果

3.1 グローバル言語の状況

国連公用語と定められている6言語のうち、中国語を除いた英語、ロシア語、フランス語、アラビア語、スペイン語の5つの言語は、地域や国家の単位を超えて使用されているグローバル言語である。これらの言語について、アジアとアフリカにおける比率を以下に示す。

表1グローバル言語の比率

	アジア	アフリカ
英語	46.2%	82.0%
ロシア語	20.6%	0.4%
フランス語	0.4%	5.8%
アラビア語	0.6%	1.9%
スペイン語	0.3%	0.9%

英語については、アジアとアフリカとの間に約36%の差異が存在する。この差異については、アジアにおいては、ロシア語の比率も大きいのに加えて、この表に上らない地域言語の比率も30.0%と高いことによって説明することができる。また、英語を公用語とする国のドメインにおいては、英語の比率はすべて80%以上である。

ロシア語とフランス語には、表からわかるように地域的な偏りが存在する。更に詳しく見てみると、アジアのロシア語ページの94.3%が、カザフスタン、キルギスタンなどの中央アジアの旧ソ連圏諸国に、アフリカのフランス語ページの90.7%が、フランス語を公用語とする25ヶ国に存在することがわかる。

アラビア語ページは、アジアでは92.7%が湾岸地域のイスラム諸国に、アフリカでは96.3%が北アフリカのイスラム諸国に存在し、それらの多くでアラビア語は公用語である。

スペイン語を公用語とする国は、アジアおよびアフリカ諸国では赤道ギニアのみである。赤

[1] Ubicrawlerに与えた収集パラメータは、以下の通りである。

ホスト探索時の最大階層数	16
1ホストからの最大ページ取得数	5,000
重複ページのチェック	行う
接続タイムアウトまでの時間	30秒

Distribution of Languages on the Asian and African domains.

†Shigeaki Kodama, Yoshiki Mikami, Chew Yew Choong. Nagaoka University of Technology.

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

道ギニアドメインでの比率は 20%と、ロシア語などに見られるような大きな偏りは存在しない。

3.2 地域言語・少数言語の状況

前節で述べたように、アジアの地域言語の全体に対する比率は、30.0%である。これに対して、アフリカの地域言語の全体に対する比率は 2.0%と低い。本節ではこのような差異が生じる原因について若干の考察を行う。

アジア諸国の公用語のうち、インドネシア語、ペルシア語、タイ語、トルコ語、ベトナム語、の 6 言語は、全体に対する比率が 1%以上で、また公用語とする国での比率が 40%を超え、第 1 位である。これらを仮に第 1 グループとする。

全体に対する比率は 1%以下であるが、国内での使用率が 10%-40%程度で、その国では情報伝達の手段としてある程度使用されていると評価できる言語を第 2 グループとすると、アジアでは、アゼルバイジャン語、ウズベク語、グルジア語、タタール語、ダリー語、モンゴル語、マレー語がここに分類される。

最後に、全体に対する比率が 1%以下で国内使用率が 10%未満である言語を第 3 グループとする。アジアでは、ヒンディー語、ウルドゥー語、ビルマ語などがここに分類される。

第 1 グループの話者人口をみると、いずれの言語も 5,000 万人を超える、地域言語としては話者数の多い言語である。これに対して第 2 グループに所属する言語は、話者数が 5,000 万人を超えない。第 3 グループに属する言語には、話者数が 5,000 万人を超えるヒンディー語などと、5,000 万人以下のネパール語などの両方が含まれている。話者人口と国内使用率に注目して分類を行うと、以下の表 2 に示すようになる。

表 2 グループ分類

		話者人口	
		5,000 万人以上	5,000 万人以下
国内使用率	40%以上	第 1 グループ	
	10-40%		第 2 グループ
	10%以下	第 3 グループ	

第 1 グループに所属する言語は、計算機上での処理のための文字コードの制定が比較的早く、フォントや入力システムなどの開発も、早かった。たとえばタイ語については 1986 年に国家規格である TIS620 が制定されている。[3]

第 2 グループに所属する言語は、グルジア語を除くと、キリル文字やラテン文字を用いて表記を行うことが可能な言語である。たとえば、アゼルバイジャン語の正書法に含まれる文字は、トルコ語の正書法に含まれる文字と同じである。

第 3 グループに所属する言語は、計算機上での処理に何らかの問題を抱えている言語が多い。たとえば、ヒンディー語にはコードに互換性のない多くのフォントが存在する。ウルドゥー語の表記は、アラビア文字のナスタリーク体という書体を用いるが、この書体の計算機上での処理は、未だに困難を抱えており、現在も開発が進められている。ビルマ語にも、フォントの混乱や、文字コードの混乱などの問題が存在する。

アフリカにおける地域言語の状況については、南アフリカ共和国の公用語であるアフリカーンス語が第 2 グループの基準を満たすのみであり、全体としては非常に低い水準にとどまっている。

アフリカの言語の計算機上での処理方法の開発の開始はアジアよりも遅く、ローカリゼーションも 1990 年代の後半になって始まったばかりである。このことがアフリカの地域言語の流通を妨げている一因であると考えられる。

4. まとめ

本発表では、アジア地域においてはいくつかの地域言語がインターネットにおける情報伝達手段として使用されている一方で、アフリカ地域においては、地域言語がそのような地位に至っていないことが明らかにした。

今後も、発表者らは、インターネット上における言語使用の実態について継続的な調査を行い、言語的デジタル・ディバイドの解消に向けた取り組みに寄与していきたいと考えている。

参考文献

[1] ITU. ICT Development Index. 2008.
[2] P. Boldi, B. Codenotti,M. Santini, and S. Vigna. “Ubicrawler: A Scalable Fully Distributed Web Crawler.” Software: Practice and Experience. 34(8): 711-726, 2004.
[3] 三上喜貴. 文字符号の歴史アジア編. 共立出版. 2002 年。

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.