母国語アーカイブ・サイトの意義と研究の進捗状況

林紘一郎・豊福晋平

▶1　問題意識

インターネットの驚異的な普及とともに，そこで使用される言語は，いつのまにか英語が主流になってしまった。それはインターネットそのものがアメリカで誕生したもので，かつアメリカのウエブサイトと，アメリカを含めた英語圏の利用者が圧倒的に多いことから，やむを得ないこととも考えられる。(1)

しかし日本をはじめ，アジアやヨーロッパ諸国などの多くの国では，必ずしも一般人が英語を自在に使いこなせる訳ではない。これらの国では，英語のできる一部の人々だけが国境を越えてインターネットを利用しており，他の人々はそうしたサークルから除外されている。やがて近未来にパソコンや通信ネットワークが一層普及してくると，そういう一部の人々と一般の人々との間にコミュニケーション・ギャップが生じ，「情報強者」と「情報弱者」に分断（デジタル・デバイド）される惧れがある。

また英語だけを国際語として認めることは，非英語圏で歴史的に形成されてきた固有の文化を同化し，衰退させる惧れもある。同じ西欧社会でも，自国文化に敏感なフランスなどは，インターネットを介してアメリカの文化が無制限に流入することに，懸念を示している。ましてや西欧社会とは文化や歴史を異にするアジアの各国では，一方で英語を公用語とすることが実利的であるとする意見があると同時に，方言も含めた自国語へのこだわりを示す向きもある。

日本語についていえば「英語を公用語とすべきだ」という提言があるとともに，これに対する根強い反論もある。実際一時ほどではないにせよ，外国人の間にも日本の文化と日本語を学ぶ人々が，多数いることも事実である。また日本語や日本文化研究のような学術交流のみならず，経済や政治の場でも日本語で情報交換できることは，相互理解を深めるだろう。

共同研究者等は，このような問題意識を潜在的に持っていたが，それをどのような視

脚注

1．近年のインターネット人口の爆発的増加に伴って，英語の比率は相対的に低下しつつあり，この指摘は当たらないのではないか，との批判もあろう。なるほど私たちの懸念は，研究を始めた1997年頃には深刻な事態であったが，状況は若干緩和されたかに見える。しかし電子メールやチャットなど，国際的交流の場合の事実上の標準語は依然として英語であり，本研究のテーマである「母国語アーカイブ・サイト」の必要性は，決して薄れていないと思われる。

点から研究すればよいかについては，初めから確たる方法論を持っていたわけではない。しかし次節で述べるような経緯で検討が深まるにつれて，他国語のサイト（とりわけ英語のサイト）を日本語で読めるようなアーカイブ・サイトを開設して，定期的に更新して常に最新情報を維持する仕組みを作り，またその逆の方向（日本語のサイトを英訳する）も可能にすれば，インターネット上の「多言語文化環境」の実現に近づくのではないか，というアイディアを共有するようになった。

▶ 2　研究の経緯

⑴　研究グループ結成の発端：1997年秋から冬

このようなわけで研究グループの結成を考えていたところ，1997年秋に偶々平成10年度の科学研究費申請に当たり，「特定領域研究(B)」において「地球有限化時代の経済社会――地球規模のインフラストラクチャーの構築」という大きなテーマで，共同研究者を募る向きがあったので，これに参加することとした。（申請代表者：佐久間章行青山学院大学教授，テーマ17「国際通信ネットワークにおける多言語文化環境に関する研究」）。この際，共同研究者に名を連ねたのは，林，西垣，石崎，宮澤の４名であった。

共同研究者に名を連ねた４名は，菅澤・豊福も誘い，研究費がいずれ承認されるとの期待のもとに，98年早々から研究会を発足させ，まず身近なインターネット翻訳ソフトの考察から研究をスタートさせた。不幸にして申請は却下されたが，共同研究者の間では，せっかく研究の方向と参加意欲が高まったのだから，他の資源を活用してもプロジェクトを起こすべきだという気運になっていった。

⑵　慶応義塾大学メディア・コミュニケーション研究所の「研究教育基金」を使った共同研究の開始：1998年当初から年度一杯

一方，本研究代表者の林は，所属する慶応義塾大学メディア・コミュニケーション研究所における個人の研究としても，この分野に集中する気持ちになっていた。なぜなら，自身のライフ・ワークである「ネットワークの法と経済的分析」の視点から見て，多様化を促すはずのマルチメディアが却って「独り勝ち」をもたらしやすいという矛盾をどう説明したら良いか，疑問を持っていたからである。

このような問題意識で97年度には同僚の菅谷教授主導のプロジェクト「メディア・コミュニケーション産業の理論的および政策的側面に関する基礎的研究」の一環として考察を始め，成果の一部を「マルチメディアとグローバル・スタンダードを考える」と題して本紀要No.48に発表した。そして98年度には独立のプロジェクト（「マルチメディアとグローバル・スタンダードに関する実証的研究」）として，最も基底的なレイアである「技術標準」と最も高位のレイアである「言語」の２チームを設け，上記論文で枠組みを設定した「レベリング」（水準を合わせていく）「ブリッジング」（異質のものを橋渡しする）の２分法を軸とした分析を行なった。

この結果前者においては，デファクト標準が中心の時代になると，その権利性をめぐった争いが起きやすいこと，アナログ時代にできた知的財産制度は新時代に適応できないこと，などが明らかになった。そこで林は，その叩き台とも言うべき視点を「ディジタル創作権の構想・序説」として本紀要No.49に発表した。

一方後者の研究からは，翻訳ソフトが相当数にのぼり，それらを評価する研究も散見されることが判明した。ついては，本格的な研究に進む前段として，翻訳ソフトの性能そのものではなく，支援ツールなどの機能を客観的に評価する手法に重点をおいて，実

証研究を継続することとした。なおこのためには，翻訳支援ソフトを購入して使ってみなければならないので，しかるべき財団に研究助成を申請することとした。

⑶ 電気通信普及財団の助成を得た研究：1999年度

幸い研究開始から実質的に３年目となる1999年度には，「インターネット利用における多言語文化環境の実現方法に関する研究」として電気通信普及財団の助成（110万円）をいただくことができたので，翻訳ソフトを実際に購入し使ってみて，その機能面を中心にした評価を行なうこととした。

多言語環境形成の一手段として機械翻訳（Machine Translation，以降MTと略）に対する需要と期待が増しているが，MTソフトの性能はまだ期待されているレベルには至っていない。MTの発展のためにはその評価が重要であるが，翻訳品質の評価の研究はよく行なわれているのに対して，翻訳機能の全般的な評価の研究は共同研究者らの知る限りまだ行なわれていない。そこで本研究では，インターネット用MTソフトの機能評価を中心として考察を行なった。対象にしたMTソフトは英日の13製品であり，1999年４月現在市販されているソフトについて，ほぼ全製品を網羅している。

研究結果の詳細は，同財団発行の「平成10年度研究調査報告書」と後出の宮澤論文を参照いただきたいが，今回の評価において各製品で機能にかなりの開きがあることが判明し，必要な機能と将来の方向性が明らかになった。本機能評価は英日版のMTソフトを対象に行なったが，これら機能評価研究は言語に依存しないので，海外のインターネットMTソフトについても当てはまる汎用性を有している。

⑷ 慶応義塾大学大型研究プロジェクトの助成を得た研究：1999年10月から2000年度一杯

共同研究者等は更に，日米の政府サイトをテストベッドとして，翻訳ソフトや要約ソフトを使って，どこまで相互理解を深められるか，実際にテストしてみたいという意欲に駆られた。そこで慶応義塾大学大型研究プロジェクトに応募したところ，幸いに研究助成（500万円）が認められたので，1999年10月以降は電気通信普及財団の助成を得た研究と平行して，「日米政府の電子化文書を使った多言語アーカイブ・サイトの開発」を実施している。なお，この段階から横山・上田が共同研究者として加わった。

具体的進め方は，日米両国政府の電子化文書（主としてウエブ・サイト）を用い，これを翻訳し要約するソフトを組合わせて「多言語アーカイブ・サイト」として維持する仕組みを開発するもので，具体的イメージは次のとおりである。

- 日米の政府が掲出している，ホーム・ページの内容から適したものを選び，翻訳ソフトや要約ソフトを使って，日英両国語で検索可能なアーカイブ・サイトを創設する。
- これが日米の政策研究者にとって、事実上のポータル･サイトとなるよう、表示方法や使い勝手などを総合的に検討し、逐次的改善を図る。
- また、これをプロトタイプとして，政府サイト以外のアプリケーションや，日英以外の多言語化へ拡張する可能性を探る。
- この過程で，ソフトの機能評価，コンテンツの適否の判定，利用者の反応，編集のノウハウ，知的財産制度のありかたなどについて，総合的・学際的に考察する。
- 研究の成果を発表するとともに、多言語文化環境への理解を深めるため，諸外国の研究者を集めたシンポジウムを開催する。

▶ 3　アーカイブ・サイトの機能と開発進捗状況

日英などメジャーな言語間での翻訳・要約の機能は，一般的に利用されているレベルのコンピュータに市販アプリケーションを導入するだけで，実現可能なレベルとなっている。翻訳品質など様々な問題は抱えているものの，それだけではあえてサーバ・クライアントのシステムを構築する意義は低い。

そこで，アーカイブ・サイトを構築するにあたっては，クライアント単独では手間がかかりすぎる部分，あるいは無駄の多い部分を代行し，運営管理者側が手間をかけなくても，自動的に内容を豊かにしていくことが可能な機能を整備することとした**（図1）**。

アーカイブ・サイトで実現させるべき主な機能は，1）定期的な相手先ページ（政府文書）の更新状況把握，2）相手先ページデータの取得，3）ページデータの翻訳・要約，といった一連の３つのプロセスを　4）ホームページアドレスや翻訳データを管理するデータベースのもとで自動化し，5）一般利用者がデータブラウズできるようなウェブ上のインタフェースを動的に提供するものである。さらに，6）各地に点在する研究協力者に対して共通のウェブ・ワークスペースを提供することで，翻訳・要約品質を確認したり，翻訳結果に一部修正を加えたり，といった作業が円滑に行なえるようになる。

試作されたホームページを**図２～４**に示す。一般利用者の利用方法としては，まずア

図１　日米政府電子化文書母国語アーカイブ・サイトの機能

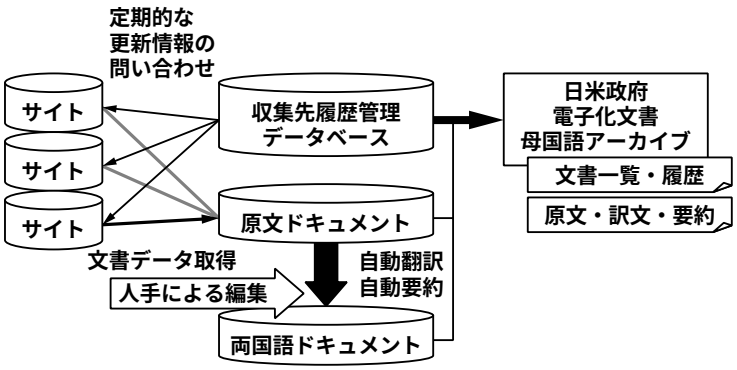

図２　アーカイブ・サイトのホームページ

図３　各文書について得られるデータの一覧

図４　原文要約データの例

Figure & Table

ーカイブ・サイトを開き（**図２**），データベースがその時点の最新情報にあわせて動的に作り出す一覧表から任意の時点の文書を選択すれば，その原典と翻訳，要約データが参照可能になる（**図３**）。画面の○×は各種文書の入手可・不可を示している。

特にアーカイブを名乗るうえで重要なのは，相手先ページデータの構造全体を定期的に取得し，これを各時点のスナップ・ショットとして保存する機能である。紙媒体のメディアとは異なり，ホームページのデータは日々変更されていくのが常であり，参照できるはずの文書が突如として消滅してしまう可能性もある。これは文書データを公開する側のサイト運営者のポリシーに負うところが大きいが，いたずらにサイト内のディレクトリ構造を変えたり，過去のデータを削除したりすると，ホームページ・ブラウザは簡単に目的のファイルに到達できなくなってしまう。

このような相手先サイトから得られる情報の不安定性や不確実性を排除するためには定期的に相手先データのコピーをアーカイブしておく方法が有効であり，もし相手先のデータが失われたり変更されたりしても，アーカイブされた内容が参照できれば，いつでも過去にさかのぼることができる。

以上の機能を実現するために，Microsoft Windows2000Serverをベースに，次のようなモジュールを利用してプロセスへの組み込みを進めている。

・収集先履歴管理データベース
・一連の処理を自動化するスクリプト（Windows Scripting Host）
・指定相手先のホームページ更新日時を取得するアプリケーション
・指定相手先のホームページデータを自動取得するアプリケーション
・翻訳または要約エンジン又はアプリケーション
・スクリプトに対応しない一部アプリケーションを画面制御するマクロツール

・ウェブ上のインタフェースを動的に提供するスクリプト（Active Server Pages）

このうち，比較的作業の進んでいる米→日のプロセスでは，相手先データの自動取得から翻訳・要約データの抽出までの自動化をほぼ完了し，現在はデータベースとウェブホームページとのインタフェース構築を主に進めている。

特にサーバ環境における自動化は，予想外の影響や結果を排除するため，なるべくコマンドライン上でスクリプト制御するのが理想であるが，Windows系アプリケーションでは制御用スクリプトに対応しない（APIを公開していない）アプリケーションも数多く，現状では画面操作をマクロ化するツールを使わざるを得ない。

データ更新状況のチェックや研究協力者向けの作業用インタフェースについては，今後の開発課題として残されている。

▶ 4　試験用サイトの選定

対象とするウエッブ・サイトは，一般の関心が高いか，行政の側が一般の国民にとくに関心を持ってもらいたいと期待しているものから選定することとし，2000年初頭にサイト選定方針として，次の各項を決定した。

・日米双方向での翻訳と要約を目指すとしても，まずは米から日への変換から着手する。
・立法・司法・行政のうち，立法と行政については専用の検索ツール（立法についてのThomas，司法についての Lexis-Nexisなど）があることから，今回は行政を主たる対象にする。[(2)]
・米国政府サイトの選定にあたっては，一般大衆向けのものと，専門家向けのものの，両方をとりあげる。
・前者の例としては，アメリカ人なら誰でもアクセスしそうな，ホワイトハウスのサイトから選定する。また，併せて麻薬撲滅運動や環境保護運動など，一部に専門用語があっても，国民一般にアピールしようとするものも含める（後段は，国際大学グローバル・コミュニケーション・センターでかって行なった，アメリカ政府サイトのデータベース化の経験に基づく助言による）。
・後者の例としては，林の専門との兼ね合いにも配慮し，FCC（連邦通信委員会）のサイトをまず検討する。

以上の結果，実験初期にとりあげるサイトとしては，次のとおりとした。

・ホワイトハウスのものとしては，時期的に見て「年頭教書」を取り上げることとした。
　http://www.whitehouse.gov/WH/SOTU00/sotu-text.html
・また専門性はあるが，一般人に分かりやすく伝える目的を持ったサイトとして，青少年犯罪や非行を扱ったサイトと，原子力を扱かったサイトの二つを取り上げる。
　http://www.ojjdp.ncjrs.org/ojstatbb/qrptnote.html
　http://www.nrc.gov/NRR/OVERSIGHT/faq.html
・FCCのサイトは一般的なものから，高度に専門的なものまで，幅広く分布しているが，今回はホットな話題のIPYVを取り上げる。
　http://www.fcc.gov/Speeches/Kennard/2000/spwek001.html

これらを対象に前３で述べたとおり，原文・自動翻訳文・対訳・母国語文などを試験用サイトに収録した。しかしその後，上記の選定方針がかなり静的なもので，一旦収録

脚注

2．本稿脱稿後アメリカ政府は，行政府のサイトへのポータルとして，http://www.FirstGov.gov/を設けた。検索機能があるほか，税務申告の用紙を取り出せるなど，「電子政府」の実現に向けての意欲と工夫が見られる。

してしまえばその後の変化に乏しいことが判明した。これは初期テスト用としては安全な対象としての価値はあったが，フェッチ・自動翻訳などが技術的に可能であることが証明された後では，試験用サイトとしての役割に欠けることになる。そこで2001年頭からは，更新頻度が激しいと思われるサイトを追加予定である。

一方，日本政府のサイトは遅れて検討することとしていたが，技術的可能性が見えてきた秋口から，対象の選定に入った。折から森内閣の「IT推進」の掛け声が轟いていたので，まずはIT関連の２省，すなわち通商産業省と郵政省にメールを送り，研究への協力を依頼した。ところがそこで起きたことは信じられない反応だった。ナシのツブテで，全くレスポンスが無かったのである。

そこで林は，それぞれの官庁に個人的ツテを頼って再アプローチし，主旨を理解していただき，翻訳された英文に手を入れる際は，アドバイスをいただくこととした。日本の官庁のこのような受身の対応の背景には，次のような問題点があるものと考えられる。

・誰がウエッブサイトの責任者か，はっきりしないし，時には外注（丸投げ）のケースもある
・英文で発表するという習慣を持たないため，翻訳された文章に責任が持てないし，責任も各局（部，課）に分散している。
・日本の著作権法では，政府文書にも著作権があることになっているため，学者の学問的営為だとしても，「ご自由に」という立場をとれない。

しかし曲がりなりにも，年明けから試験用サイトを選定し，実験を開始する予定である。

▶ 5　問題点と今後の展望

問題点として，既に判明している諸点は，次のとおりである。

・翻訳・要約ソフトの品質レベルが高くないので，人手による作業が不可避で相当数にのぼる。
・分野によっては専門用語が難しく，ドキュメント作成者との協同作業が必要である。
・翻訳結果を，原文のテキスト情報以外の部分と関連付けながら配置するのは，意外に難しい。
・ウエッブ・サイトの構造によっては，本システムが有効に機能しない場合がある。
・アメリカ政府のドキュメントは著作権フリーであるが，日本政府のものは明確でない。
・日本政府のウエッブ・サイトには，一部英語化されたものがあり本システムとの調和が必要である。

とりわけ日本政府のものを英訳する場合には，著作権の問題のほか所謂官庁用語などの壁もあり，未だ一部熱心な官庁と個別に対応している段階で，幅広い選定は今後の課題である。来年度以降はこのシステムを研究メンバー以外にも限定的に開放し，実用化に向けた実験をしてみることにより，原因の解明を行い解決策を探っていきたい。

上記問題点のうち最初の２点は，辞書の充実などで精度は飛躍的に向上するものと思われる。次の２点は純技術的問題であり，創意工夫により改善が期待できる。むしろ問題は，最後の２点のような技術以外の要素であり，その解決には困難を伴うことが想定される。

ところで，これまでに開発しメンバー間で試験したものは，プロトタイプにすぎず，このままで大規模な利用や商品化に耐えられるものではない。そこで我々は研究をもう１歩進め，対象とするサイトを増やすと同時に，実験参加者も拡大した実証実験を実施したいと考えている。

上に述べたように，解決すべき問題点は技術上のものというより，社会的なものである。実証実験を行ないたいというのも，主としてその側面からのニーズに基づいており，この実験は技術実験であると同時に社会実験の要素を持っているので，「人と機械の調和」を主眼とする財団などに助成金を申請し，研究の継続と拡大を訴えているところである。

●謝　辞

アーカイブ・サイトの立ち上げと改善には，花川憲司君（慶応義塾大学大学院政策・メディア研究科修士課程）の全面的な協力を得たので，ここに記して謝意を表します。

●参 考 文 献

加藤弘一（2000）電脳社会の日本語，文春新書094，文芸春秋
鈴木孝夫（1999）日本人はなぜ英語ができないか，岩波新書622，岩波書店

（林紘一郎　慶応義塾大学メディア・コミュニケーション研究所教授）
（豊福晋平　国際大学グローバル・コミュニケーション・センター講師）