ニューガラス関連学会

「言語天文台（三上喜貴先生）を聴いて」

セントラル硝子株式会社　ファイン硝子事業企画室

冨岡　孝夫

Report on Memorial Lecture of NGF General Meeting

Takao Tomioka

Fine Glass Business Planning Department,Central Glass Co.,Ltd.

2006年6月2日（金）に社団法人ニューガラスフォーラム第19回通常総会記念講演として，三上喜貴教授（長岡技術科学大学　経営情報系）による「言語天文台」と題した90分の講演が虎ノ門パストラル新館ロゼの間で行われた。本報告では先生の講演内容を簡単に紹介する。

講演の冒頭，本件は科学技術振興機構（ＪＳＴ）が実施している社会技術研究プログラム（研究領域「社会システム／社会技術論」）として採択され研究している旨の報告があった。

「言語天文台」名称の由来

世界には6,000前後の言語が存在しており，この数はたまたま我々が見ることができる星の数にほぼ等しいといわれている。パラボラアンテナを使って星の詳細情報を集めるようにインターネット上で使われている世界の言語を集め分析するプロジェクトなので言語と天文台をかけて「言語天文台」と命名した。ちなみに「言語天文台」はすでに商標登録を済ませている。

〒101-0054　東京都千代田区神田錦町 3－7－1
TEL　03-3259-7357
FAX　03-3259-7888
E-mail : ttomioka@cgco.co.jp

言語の多様性

パリ，イハンビクタス広場にある「ジュテームの壁」は長さ10m，高さ4mにわたるもので世界中から集められた【愛している】という言葉が各々の国の言語で刻みこまれているというもので言語や愛の言葉を通じて平和のつなぎとなってほしいというメッセージが込められている。

ＳＩＬ（Summer Institute of Linguistics）の説では手話，アイヌ語等もカウントすると世界には6000前後の言語があるが，長い歴史の中では共通化されたりして減ってきている。このように多数の言語がある中で一番多く翻訳されているのは聖書であり，部分訳も含めると約2,200言語に翻訳されている。2番目は世界人権宣言であるがweb上ではエンコード化されていない言語はスキャナーで取りこみそのまま貼りつけられている。ちなみに日本語関係で一番多くの言語に翻訳されているのはＮＨＫで放送された「おしん」で「窓ぎわのトットちゃん」も結構多くの言語に翻訳されている。

文字の多様性

中西印刷の先代社長の中西亮氏が世界中を回って集めたコレクションをもとに作った世界文

字地図というものがあり必ずしも現在と一致していないものもあるが大別するとラテン文字圏，キリル文字圏，アラビア文字圏に分類することができる。

また文字は宗教と密接な関係があり，主なものを挙げると漢字【大乗仏教】，ラテンアルファベット【キリスト教】，キリル文字【ギリシャ正教】，アラビア文字【イスラム教】，インド系文字【当初ヒンズー教，後に上座部仏教】，チベット文字【ラマ教仏教】，ヘブライ文字【ユダヤ教】となる。

文字は漢字に代表される表意文字とアルファベット，ひらがなに代表される表音文字に分類することができる。表意文字は語彙数としては 10^4〜10^5 と非常に多いもののコンピュータ処理上は比較的やりやすく，逆に難しいのは表音文字で中でも結合音節文字に分類されるインド系文字はコンピュータ泣かせである。

デジタルデバイドと言語・文字

世界の所得水準毎での固定電話，携帯電話，インターネットの普及率を 6 年前と比較してみると中国，インド圏の所得レベルにおいても携帯電話が固定電話を上回っているがインターネットの普及率はまだまだである。世界の人口 65 億のうち 4 割はラテン文字圏でありインターネット人口の 8 割以上がラテン文字圏となっている。ちなみに活字文化の普及率の目安となる印刷用紙一人あたりの年間使用枚数を日本と北朝鮮で比較してみるとＡ 4 換算で日本は 2 万枚であるのに対して北朝鮮はわずか 3 枚となっている。

また検索エンジンの代表格である google もすべての言語に対応しているわけでないので自分の国の言語で検索できるのも限られている。

情報技術とローカライゼーション

15 世紀から活字印刷が始まり，いまではほとんど見られなくなったタイプライターが全盛であった時期もあった。またハングル文字をタイプライターで打つ場合は重ね打ちするなどの工夫もなされていた。活字印刷，タイプライター，電信，コンピュータというように文字の印刷方法が進化してきたが，いずれの時代の情報技術も何らかのグローバルな技術標準が不可欠であり，言語に応じたローカライズが必要である。

言語天文台の目的と仕組み

ネットワーク上で各言語がどのように使われているか分析するツールをつくることが目的でまずは言語判定エンジンを作り約 300 言語までは判定できるようになった。

研究終了までにカントリーコードをもとに何語のページが何ページあったかというレポートと文字コード別利用状況のレポートを作成する予定。

我々が肉眼で見ることができる星は六等星まででそのうち一等星が 21，二等星が 67 である。言語天文台の観点からいうと google が 35 言語に対応，XP が 71 言語に対応しており，ほぼ星の明るさの分類と相関を持たせることができることもネーミングの由来となっている。

実際の仕組みとしては，インターネット上から自分でデーターを持ってきて言語を判定することになるが各々のホームページにあるリンクを次から次へと辿るという地道な作業でイタリアのミラノ大学が開発したソフトを使用している。

現在，長岡技術科学大学には 40 台のサーバーがあり合計で 10 T バイトの容量があるが世界中には 80 億ページ以上があり 1 日に持ってこられるのは 500 万ページ程度とまだまだ容量不足なので，まずはカントリードメインのみに限ってデーターを集めている段階である。

集めたデーターの URL を分析したり，タグ情報，テキスト，サーバーレスポンスの解析をおこなうことで，人口当たりのページ数，言語別ページ数，自国語ページ比率，通信インフラの良否等を調べることができる。

アフリカ調査の結果

手始めにまずアフリカの国のドメイン名から国別人口とページ数についてまとめてみると人口の少ない Ascension (AC), Seychelles (SC) のページ数が多いことがわかり,よく調べてみると AC は Academic の略称でもあるので大学関係がドメインを買っていることがわかった。また SC はオーストリアの株式会社の略称と同じなので同様にオーストリアの企業が買っているため，見かけ上多くなっているようである。さらによく調べてみると空のページも多く存在していることもわかった。これはサーチエンジンをだますスパムページで google に代表されるサーチエンジンはリンクの多いもの程上位にランキングするのでわざとダミーリンクのページを外に多く作っているようである。またアフリカではサーバーの物理的所在地も 8 割が国外となっており，その国のドメイン名がついていても母国語で表示されていないものが大多数となっているのが現状である。

以上，先生の講演内容をまとめたつもりである。個人的には天文に興味があるのでいったいどのような内容なのかと興味津々で拝聴させていただいた。偶然かもしれないが，星の等級と現在使われている言語との関係が比較的似ているのでネーミング自体はそれ程違和感ないと感じた。また先生の貴重なコレクションであるめずらしい外国の紙幣も回覧していただき，少しは知らない国の言語にも直接触れる機会が得られた。先生の講演の主旨を十分伝えられていない部分があるかもしれないがご容赦いただければ幸いである。