デジタル文献学の試み

書棚を整理していると古い本に挟まった幅20センチほどの黄土色のカードが見つかった。これはパソコンが普及していなかった４０年前、大学の電算機室のコンピュータで中世スペイン最古の武勲詩『わがシッドの歌』全編を分析したときに使ったパンチカードと呼ばれていたものである。一面に小さな穴が穿孔されたカードには80文字だけの情報しか載せることができなかった。校正の度にカードを捨て、最終的に3000枚以上の資料体（コーパス）を完成させた。これをカードリーダで入力し、やはりパンチカードに穿孔したプログラムで処理すると、その結果を大型のラインプリンターが新聞紙幅ほどの連続用紙に騒音をたてて出力した。せっかく作った資料であったが、段ボール箱に詰めたパンチカードは、その後、紙テープ、磁気テープ、フロッピーディスク、ハードディスクなどの媒体にコピーし、現在はクラウド上に置いてある。当時のカードはすべて廃棄したのだが、栞の代わりにした一枚だけが手元に残った。

あの頃からコンピュータを使った「デジタル文献学」の方法を探ってきた。さまざまな開発言語を使ってみたが、最近ではMicrosoft OfficeのVBAを中心にしている。文系の学生が履修している私の授業では高度な技法を使うことは困難なので、コーディングが比較的簡便なVBAを選択した。どの言語の文献もユニコードによって簡単に処理できる。履修者はOfficeのオブジェクトが簡単なコードで操作できるので面白い、という。彼らと一緒に私も毎日のように新しいことを学んでいる。

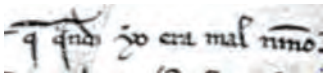
大量なデータで小さな発見

スペイン語には「エニェ」 ñという、 nの上に波線をつけた文字がある。たとえば「スペイン語」はespañolと書き、「エスパニョル」と発音するが、この中にもñが使われている。この文字の起源については、その母体であったラテン語にはなかった音なので、中世スペイン語の古文献を探らなければならない。そのようなときデジタル化した文献資料が役立つ。たった一つの作品だけでは信頼できるデータにならないが、現在ではスペインおよびラテンアメリカの各地の大学の研究グループが次々に大規模な資料体を完成させインターネットを介して無償で提供してくれるのでありがたい。そのようなグループの1つに13世紀から17世紀に渡ってスペイン各地で発行された公証文書資料を大量に収集している機関がある。その資料の中で「エニェ」の出現をたどると13世紀の略記法に遡ることがわかった。当時の書記たちはnb, nc, ndなどのnを省略し、その代償として単語の上に波線を引いていたのだが、これを略さないこともあった。ところがñが由来するnnのときだけは14世紀前半から一貫して略記し、完全形にすることがなかった。これは特有の子音を表示していたためであろう。このときにñの文字が誕生したと考えられる。このような小さな知見でも各時代を網羅する大量の資料がなければ

得られない。

ワンポイントの調査

中世ラテン語では、語頭にv、語中にuというように位置による文字の使い分けをしていた。これはイギリスの『ジェームズ王聖書』でも同様であったことを言語情報科学専攻の同僚の先生からご教示いただいた。そして18世紀のスペイン語で現在のようにvで子音を、uで母音を表記するようになった。中世以来のラテン語法から近代の新規範に移行する過渡期にあった16-17世紀ではu-vの文字が混用されていたというのが従来の通説である。同じ単語であってもuで書かれたりvで書かれたりしていたからである。しかし、手稿本ならばそのような混用があったかもしれないが、念入りに活字を組んでいた印刷本ではどうだったのだろうか？

(a)  (b)  (c) 

(a) q<ue> q<ua>ndo yo era mas nin<n>o (Sánchez-Prieto, 1995, Textos para la historia del español. t. II. Archivo Municipal de Guadalajara. Universidad de Alcalá de Henares (6) Sevilla, 1251); nn: (b) dona YOLA<n>T mi mugier (Córdoba, 1260); (c) en el anno [id. (4)]

大学の文学の授業などで扱う古文献の多くは文献学者が考証を与えた校訂版が使われているが、言語史の研究では手稿本や初版本の写真を使わなければならない（上図）。そこで試みにスペイン文学史に必ず登場する有名な本を6冊選んで、スペイン国立図書館の複写を見ながら、そのデジタル資料を作成してみた。これを自作の分析器にかけて文字の分布を調べてみると、それぞれの作品で、旧式（中世ラテン語式）と新式（近代の規範）のどちらかのu-v使用法が統一的に使われていることがわかった。なるほど、すべての作品を一緒にして観察すれば文字の混用があったように見えるが、それぞれの印刷本には一定の規則があった、ということである。このように時代、地域、メディアを限ったワンポイントの調査であれば、比較的小規模の資料でもどうにか実行できる。

行列

これまでの研究では、先のñやu-vのような一定の形式に焦点をあてた分析法を用いていたが、数年前から全体の流れを総合的に観察する、という方針をとっている。そのためには、時間と空間を深く広く渉猟できる大規模な資料が必要である。個人でできることには限りがあるので現在では国際的な研究チームが組織され、メンバーたちは頻繁に連絡をしながら資料の充実と新しい研究方法の開拓に勤しんでいる。私たちとしては、言語形式と年代・地域を多次元・多変量の行列にして、線形代数を応用した行列計算をしながら、全体のパタンを観察する、という方法を提案している。多変量解析の利点は、はじめからデータ行列を一定の基準で先に分類してから分析するのではなく（前範疇化）、むしろ、原データを分類せずにそのまま分析し、その結果を解釈しながら後で合理的に分類できる点にある（後範疇化）。私たち

が「集中分析」と呼んでいる後者の方法によって、前範疇化法では見えなかったことが全体的に把握できるようになった。

イベリア半島の北東部および地中海の島嶼部ではスペイン語とは姉妹関係にある独自の言語カタルニア語が話されている。その動詞の変化形には人称・数の異なる多様な形式が使われる。バルセロナ大学の研究者が、私たちの方法に関心を示し共同研究を提案した。次の図は縦軸に動詞の人称・数、横軸に149の調査点を置き、それぞれの形式の有無をプロットしたものである。それぞれの語形が連続する地点に分布していることがわかる。

この図の縦軸と横軸の項目の順番を、集中分析法の数値計算によって変え、反応点が左上から右下に順次移項するように変形させると、次の図のようになる。

このようにして再配置された反応点は、その縦軸と横軸の並び方に一定の解釈を与えてくれる。この場合は、イベリア半島北東部→東部→バレアレス諸島→サルジニア島の一部、という連続線が形成されている。この方法によって、従来の言語地理学で提示されていた単純な「等語線」では困難であった複雑な分布の表示と解釈が可能になった。

歴史的変化と地理的変異

現在、私たちは中世、近代、現代のスペイン語の千年間の時代推移と、ヨーロッパ、アフリカ、南北アメリカ大陸にかけて分布する広域スペイン語の地理的変異をいくつかの分析手法を使いながら総合的に把握する方法を模索している。そのとき、言語資料と分析装置を自作すればその研究について説明が可能になる。外国のグループと連携しながらも、それに完全に依存するのではなく、自分たちの独自性も発揮したい。そうすれば互いに裨益することが多いのではないか、と考えている。

次に作成する資料は私が学部生のときに電算機室で手がけた『わがシッドの歌』の校訂版である。パンチカードを使って作成した当時の資料と分析装置の媒体には両腕に抱え込むほどの質量があった。現在のものはディスプレイに展開する文字列を目視で追うだけの、触感のないコーパスとコードとなった。プログラムのコーディングは頭が熱くなるほど創造的な仕事であり、苦しいものの熱中することが多い。一方、言語コーパスの作成は資料を見ながら手で入力するだけの非創造的な作業であるが、これもまた不思議に楽しい。

上田　博人（うえだ　ひろと）

［**生年月**］1951 年 10 月
［**出身大学又は最終学歴**］スペイン、アルカラ大学
［**専門領域**］デジタル文献学、スペイン語学、中世スペイン語
［**主たる著書・論文**］**(3 本まで、タイトル・発行誌名あるいは発行機関名)**
『スペイン語文法ハンドブック』(研究社 , 2011)、『プエルタ新スペイン語辞典』(研究社 , 2006)、"Palatal graphemes in a medieval Spanish biblical text: A corpus analysis of ≪ i, j, y ≫ in Genesis, Biblia de Alba", Corpus Analysis and Variation in Linguistics, edited by Yuji Kawaguchi, Makoto Minegishi and Jacques Durand, John Benjamins Publishing Company, pp. 239-257.(2009)
［**所属**］情報学環文化人間情報学コース、総合文化研究科（言語情報科学）
［**所属学会**］日本イスパニア学会、計量国語学会、コーパス英語学会、ラテンアメリカ言語学文献学会