計量国語学 30 巻 4 号（2016 年 3 月）pp.216-233.

解説

データの視覚化（4）
―Excel 散布図のグラフ・地図への応用―

井上　史雄（国立国語研究所）

要旨

本稿では視覚化によって言語研究を推進する発想と技法について論じる．数値データをグラフや地図によって目に見える形にすることによって，データの構造が読み取りやすくなり，かつ新しい発想が生まれることがある．多変量解析法を適用することによってデータの内部構造が明らかになるが，いったん構造が分かれば，もっと単純な集計でも内部構造を示しうる．理論や仮説を検証するための研究もあるが，得られたデータそのものに語らせる研究も必要であり，それには分析技法と結果の表示（視覚化）技法が有用である．

キーワード：Excel，散布図，ラベル付け，XY Chart Labels

1．岡崎敬語における数表と散布図の得失

1.1　本稿の目標

シリーズとしての「データの視覚化（2）」（林 2015）で，エクセル散布図作成の基本技法についての解説があった．本稿ではそれを踏まえて，実際の言語データへの適用例を紹介する．散布図を活用すると，折れ線グラフで表現できないことを表わせるだけでなく，変数間の相互関係を線で結んだり離したりして，表示できる．また地理的位置も忠実に表現できて，方言地図も作れる．本稿では既出の資料をもとに，視覚化について，少しの加工でもっと多くが読み取れることを，実例で示したい．

通常の折れ線グラフは，作図上，散布図の一種と見なすことができる[1]．ちょうど正方形が長方形（さらには台形，四辺形）の一種であるのと同様である．折れ線グラフの横軸が等間隔という制約がエクセル散布図にはないので，多様な応用が可能である．

狭義の「散布図」は後掲の図 6 のように 2 軸の値によって点がプロットされた（だけの）ものをさす．しかしエクセルの「散布図」では，図 2，図 3，図 7 のように関係の深い点を線で結ぶことができる．また図 8 のように，点に様々な記号を付けることもできる．本稿ではこれらを広義の散布図として，エクセルの散布図機能を使ったものを紹介する．敬語にも適用した（井上 2011）．

1　エクセル散布図の横軸を等間隔にして，線でつなぐと折れ線グラフが描ける．

1.2　岡崎敬語データの性格

まず数値の表示技法にふれる．エクセル散布図を上手に使うと，従来にない発想に結びつくことがある．その効果的な例をあげる．愛知県岡崎市では，国立国語研究所が 3 回にわたって敬語の調査を行ったが，調査年は切りのいい年ではない．

表 1 に数表部分を示す（デフォルトでは，小数点以下の 0 は表示されない）．この数字だけから増加の傾向を適切に読み取ることは難しい．

表 1　岡崎敬語調査での「ていただく」使用率平均点

	1 次	2 次	3 次
調査年	1953	1972	2008
岡崎ていただく平均点	2.420408	2.835	4.199346

1.3　調査年代の表示

図 1 には 3 回の岡崎敬語調査での「ていただく」使用率平均点を，折れ線グラフによって，等間隔で示した．線の傾きからみて，第 2 次から第 3 次調査にかけて，大幅な増大があったと読み取れる．

作図の手順は次のようである．表 1 を囲み，エクセルのメニューバーの「挿入」の「折れ線」で「2D 折れ線」の一つを選んだ．

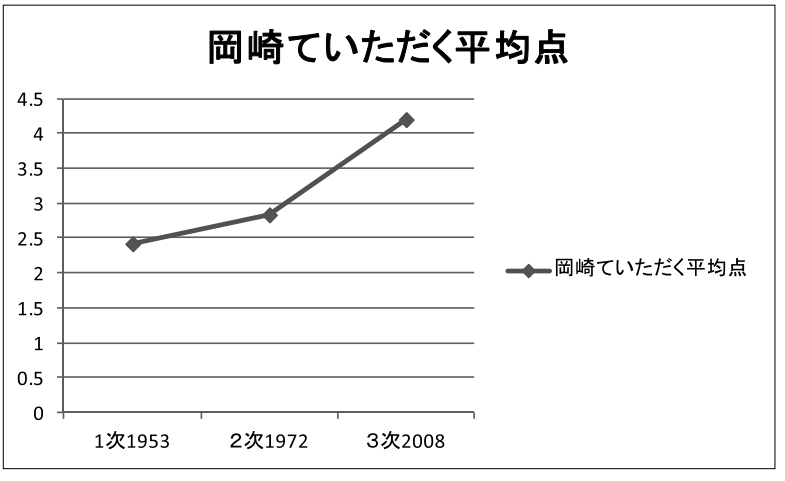

図 1　岡崎敬語調査の「ていただく」使用率平均点：全 3 回等間隔

1.4　エクセル散布図による調査年代表示

しかし岡崎では第 1 次から第 2 次，さらに第 3 次まで，調査間隔に大差がある．このような場合に，エクセル散布図を使うと，調査間隔を実年代に忠実に表わせる．図 2 に示す．第 1 次から第 2 次，第 3 次調査にかけて，ほぼ同程度の増大があったと読み取れる．

手順としては，同じく表 1 で，エクセルの「挿入」の「散布図」の一つを選んだ．横軸の間隔が 19 年，36 年になった．なお軸の表示を変える手法もある．この場合はグラフ横

軸を右クリックして，「軸の書式」をクリックし，出てきた枠の中の左欄，「表示形式」を選ぶと，「標準・数値・通貨」など以外に「日付・時刻・・・その他・ユーザー設定」などが選択できる[2].

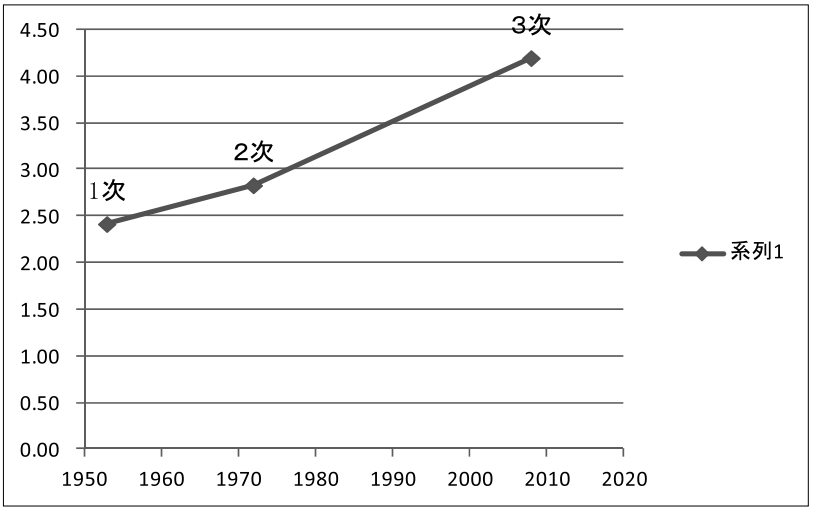

図 2　岡崎敬語調査の「ていただく」使用率平均点：調査実年代

1.5　散布図への近似直線挿入

図 2 に近似直線を入れてみる．もし直線的に過去から増大したと考えたら，いつから普及したかの見当がつく．図 3 に示す．横軸は調整できる．横軸をクリックして，「軸の書式設定」を選び，「軸のオプション」の「最小値」を 1840 にした．近似直線の下端から見て，1880 年前後に使用開始と思われる．ただし普及の S 字カーブを描くとしたら（井上 2000，横山・真田 2007），もっと以前から使われていたことになる．

図 4 に，以上で用いた愛知県岡崎市データの一部の画像を掲げる[3]．後述（3.4）の近似直線の数式が入っている．

1.6　他の岡崎敬語データへの適用

岡崎第 2 次調査は 19 年後，第 3 次調査は 36 年後なので，19 年か 18 年の間隔で年齢層を区切ると，3 次にわたって，同一年齢層の数値を示すことができる．西尾（2015）は年齢層を 19 年きざみにした．ちょうど山形県鶴岡市の共通語化調査について，米田（1997）が 11 年きざみの年齢層を採用したのと同様である．鶴岡では調査間隔が 4 回とも共通になり，しかも少人数のセルがなくなるという利点があった．

2　図 3 のグラフは，プレゼン用に別途作図したので，ラベルなどに図 1 の折れ線グラフと食い違う点がある．林（2015）の解説による技法で付加できる．図 4 も同様．
3　誌面では（薄い）黒になるが，実際の作業では青になるので，そのまま提示する．

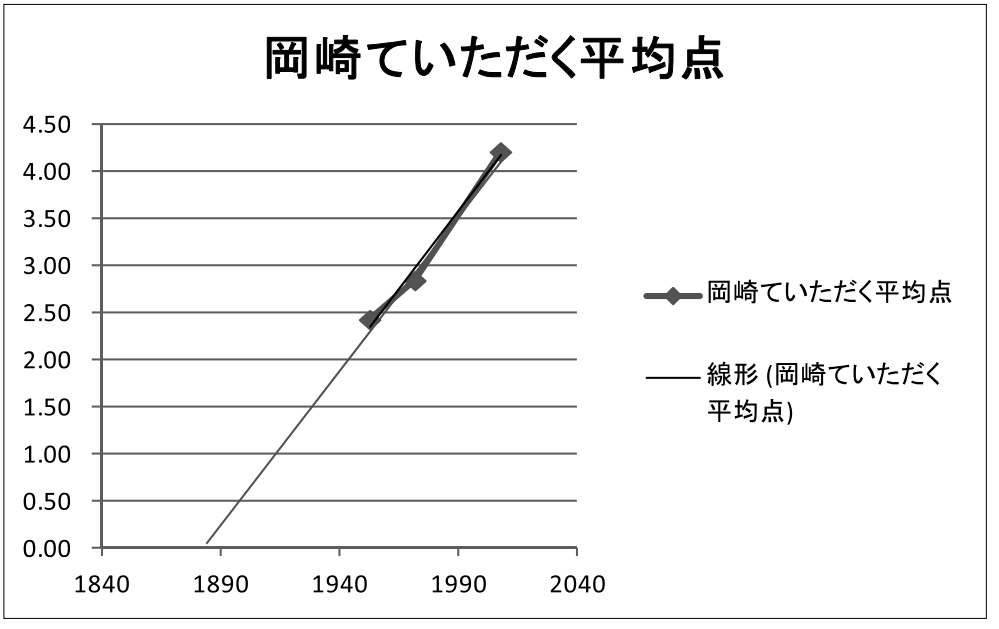

図 3　岡崎敬語調査の「ていただく」使用率平均点：調査実年代と回帰直線

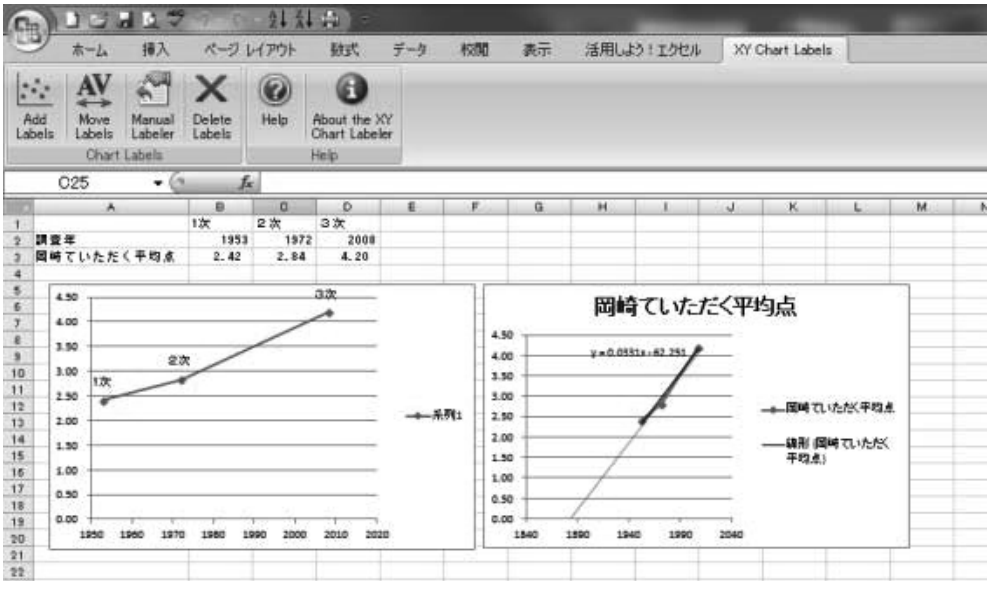

図 4　岡崎敬語調査の「ていただく」散布図の手順

岡崎敬語調査の分析では，エクセル散布図の機能を様々に活用した．多変量解析法の結果提示にも使った．3 回の調査結果を 10 年きざみの年齢層に分けて，生年実年代に忠実に図化した（井上他 2012）．また，パネル（同一話者追跡）調査の提示にも使い，個人が

どの程度の変化を起こしたかを，全体として表示できた．ウェブで多様な適用例を公開してある．「大規模経年調査資料集」参照（http://innowayf.net/）．

2．外来語の時代間隔表示における散布図利用

2.1　外来語の時代間隔表示

以下では散布図利用の効果を示す例として，外来語の時代別表示を取り上げる．荻野（1988）のグラフでは，普通の折れ線グラフを使ったために，時代間隔が均等で，徳川 300 年も大正の 15 年も等間隔で表示されている．散布図を使えば，各時代の実際の長さに対応したグラフができる．ただし各時代の区切りの年にグラフのマーカーを置くと，時代の長さが忠実に反映されないので，各時代の真ん中の年を計算した．

図 5 に結果を掲げた．見て失敗と思った．明治以降がこみあって，線が読み取れない．荻野（1988）の元のグラフのほうが，近代の中の違いがよく読み取れる．しかしグラフ全体を拡大して，線の間隔を広げた．また入れる言語を取捨した（ギリシャ，ラテンをグラフから削除し，中国語を入れ，総計の数値を入れた）．グラフがやや見やすくなった．

縦軸は語数を示し，荻野（1988）と同じく対数表示である．線の色と形は，エクセルの余計なお世話である．全部を黒にし，言語グループごとに似た線にした（林 2015）．言語名略称のラベルは「挿入」「テキスト」で，手で追加した．グラフ下端に各時代名と区切りの年を「挿入」「テキスト」の機能を使って入れた．

2.2　外来語全体視覚化の効果

数値化し，グラフで視覚化したからこそ読み取れることがある．このグラフの読み取りでもっとも効果が大きかったのは，外来語の総計 TOTAL を入れたことである．線とマーカーを――●――に変えて（手法は林 2015 参照），際立たせた．外来語の全体像が提示されると，別の印象が得られる．そもそも長期にわたる言語変化がどんなパターンを描くかという理論的見通しと関係づけうる．総計を太い線にしたところ，大まかには，対数曲線に従って外来語全体が増加するかに見えた．倍々ゲーム，幾何級数的増加，加速度的増加と見られる．算術級数と違って，対数曲線による表示がふさわしい．

こう考えると，鎌倉時代の凹みは，戦乱の時期，武士の時代のためとみられる．平安時代の遣唐使廃止による国風文化の隆盛のあとであり，外に開かれた時代ではない[4]．

江戸時代の鎖国は，全時代を通してみると，外来語に影響が少ない．長崎を経て実際に物資の流入があり，品物が名前とともに流入したおかげである．開港（日米和親条約 1854）以降の 1 世紀半は，語数全体で見ても直線的増加にならない．戦時中の鬼畜米英，敵性語排斥の風潮にもよる．

4　なお蒙古襲来のころの 13 世紀には文献そのものが少ない．国語史概説書にあるとおりで，国語学辞典所載の文献数も減る（世紀別に入力，グラフ化した未公表グラフがある）．日本国語大辞典第 2 版でも，世紀別，年単位に検証できる．なお LAJ（日本言語地図）標準語形初出年にも偏りがあり，中世初出の語は少なく，室町時代に入って，キリシタン資料で多くなる（井上 2001）．

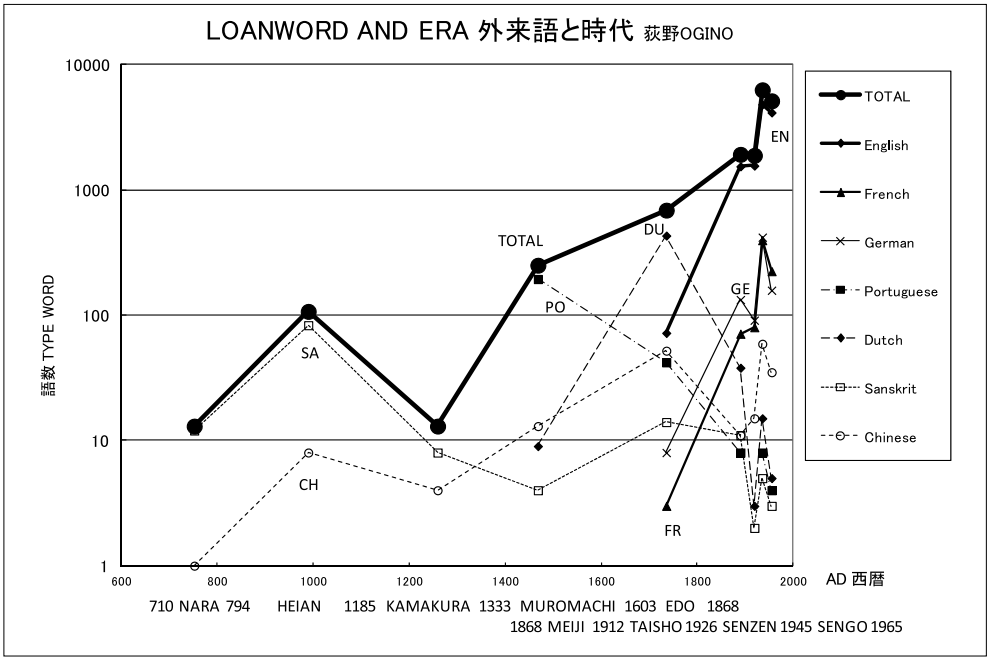

図 5　外来語の流入時期と原語　荻野（1988）

2.3　個別言語の考察

以上では荻野（1988）で扱われなかった総計 TOTAL の数値をみた．以下では個別言語をみる．中国語 CH は，グラフには入っていなかった．今回図 5 に……○……として入れたところ，総計と似た増加傾向を示した．平安時代から江戸時代にかけて，鎌倉時代を除き（対数表示では）順調に増えたように見える[5]．対外（対中）関係の歴史的推移と対応する点では，西洋の諸言語と同様である．サンスクリットも中国語経由で流入した．同じサンスクリット起源でも，漢訳仏典のような中国語への翻訳・意訳方式を経たものや，和語漢語への翻訳・置き換えもあったはずである．辞書などでの「外来語」としての扱いの限界がある．

これ以外の（印欧諸語の）推移は荻野（1988）で考察ずみである．図 5 のグラフでは，時代間隔を忠実に示したために，室町時代以来の西洋の諸言語の勢力交代，幕末開港以来の英仏独語の流入がさらに顕著に観察された．ことに英語の一人勝ちが目立つ．なおどの言語でも（かつ総計 TOTAL でも）「現代」（＝戦後）の数値が少ないが，これは元になった楳垣（1972）の発行（編集）時期が昭和（戦前）から年数が経っていなかったためである．平成に切り替わるまでの数字に変えると，十分な年数が確保され，もっときれいなグラフができる可能性がある．

日本語には Google Ngram Viewer の適用がないが（井上 2014a），日本国語大辞典のデ

5　鎌倉時代のいわゆる唐宋音による借用語が，語種として「漢語」に入れられることが，「外来語」の総数の少なさに影響している可能性がある．

ータを利用できれば，または歴史コーパスが拡充されれば，外来語の長期的変動が明らかになるだろう．

3．外行語の多変量解析：グラフ化とラベル付加

3.1　多変量解析結果の散布図表示

データの視覚化が計量言語学的研究に役立つ典型は，多変量解析結果の読み取りである．因子分析他では因子負荷量（因子係数）や因子得点が数表として出力される．数値の一覧表では，相互の値の離れ方，固まり方が分かりにくいが，柱一本に軸一つ分の数表目盛りを付けるだけで，全体の分布が分かる．さらに二つの値のクロスで散布図を作ると，2 次元の平面に位置づけられ，要素（変数）の塊が把握しやすくなる．

因子分析では，varimax 回転を施すなど様々な手法の選択が可能で，結果の比較が簡単にできる．主成分分析はじめ他の多変量解析でも，結果の表示には二つの値の 2 次元表示で深い読み取りができる．林の数量化，対応分析（コレスポンデンスアナリシス）でも同様である．

手書きグラフの時代の苦労は，今はなくなった．自動的にまたは付属オプションの選択でグラフが出る（ただし SPSS では，因子得点の表が分かりにくい形で出力される）．結果をよりよく解釈するためには，散布図に適切なラベルや補助情報を与えるほうがいい．

3.2　ラベル付加プログラム XY Chart Labeler の効果

多変量解析の結果のグラフ化では，数値をテキストファイルとして出力したあと，他の統計ソフトでグラフ化するほうがきれいに作図できることがある．文系のデータでは個々の点が意味を持つことがあるが，エクセルではプロットされた個々の点にはラベルが付かない．従来作成した散布図では，線や点の数がそれほど多くなかったので，ラベルを付ける手間もかからず，手で入れるので十分だった．しかしデータ数が多いと面倒である．

XY Chart Labeler はラベル付けに便利である[6]．ウェブで「XY Chart Labeler」を検索し，ダウンロードしてインストールすると，使っていたパソコンのエクセルにこのソフトが機能追加される．「エクセル」の「ホーム」のメニューバーに「XY Chart Labels」が追加表示される（使い方については 3.3 および 4.4 参照）．

6　XY Chart Labels は AppsPro が提供する，フリーのエクセルのアドインソフトで，名称を「XY Chart Labeler」という．2015 年 12 月現在のリンク先は　http://www.appspro.com/Utilities/Utilities.htm　である．exe 型の導入用ソフト「XYChartLabeler.exe」をダウンロードし，ダブルクリックするとインストールが開始される．（編集委員による補遺）

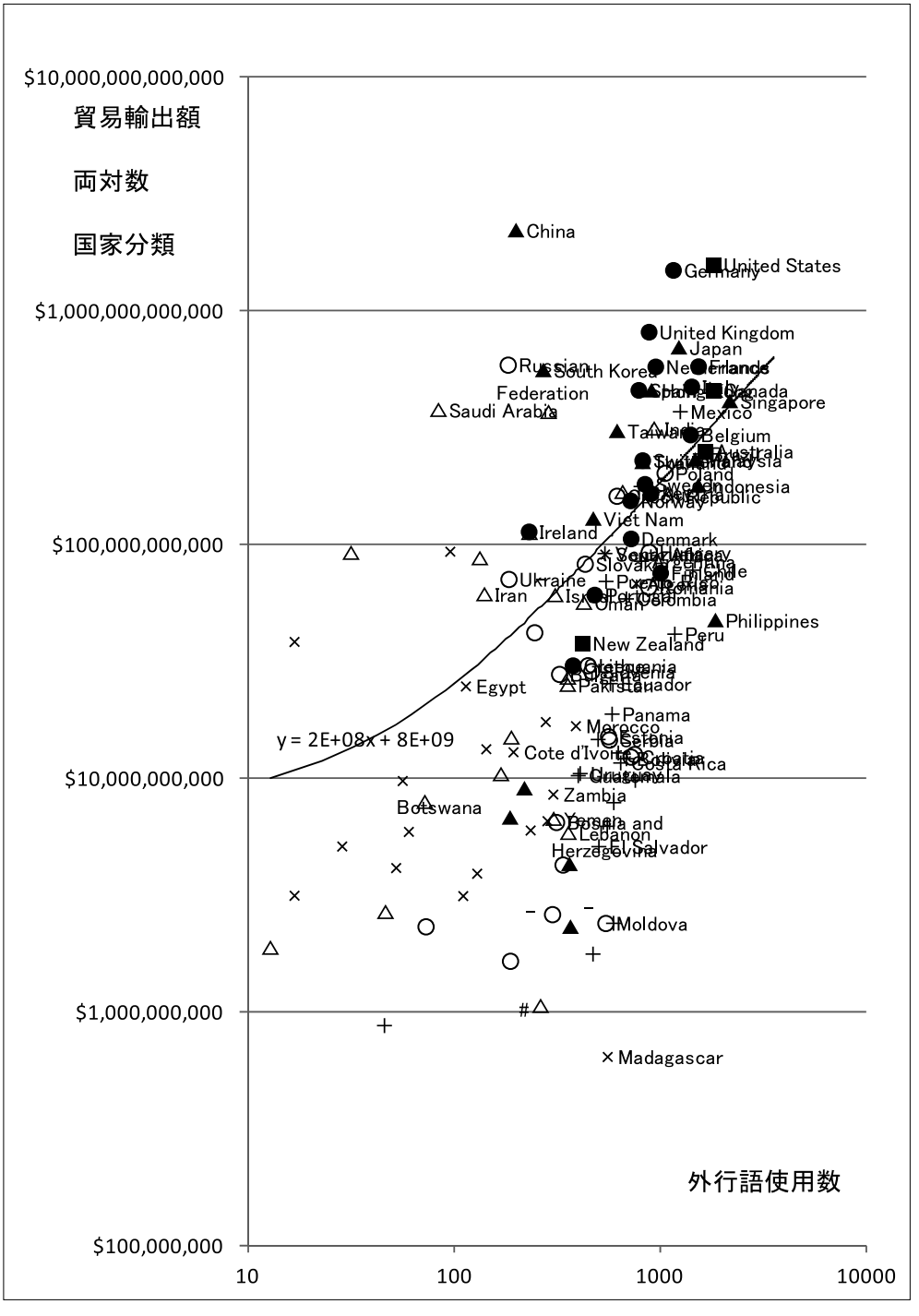

図 6　外行語使用数と貿易額の散布図

3.3　XY Chart Labels のラベル機能の使い方

以下に，XY Chart Labels の使用法（何ができるか）を解説する．作成中のグラフを開いた状態でクリックし，ラベルの入力してあるセルをドラッグする．あらかじめ隣り合う行（または列）にラベルにふさわしい文字列を入力しておく必要がある．

井上・柳村（2015）の外行語[7]の研究について，投稿後，ラベル機能を使って図の改訂を試みた．国名だけをラベルにしていたが，国名の前に言語分類に関する記号を入れた．初校での直しが不可能だったので，ここに提示する．

単純な情報付加により，図 6 の傾向が簡単にかつ明確に読み取れるようになった．個々の国名を見てどの地域かを判断する手間が省ける．日本語からの外行語が多いと分かった地域（西欧と東アジア）を黒塗りで目立たせたので，図の右上が黒っぽく見える．

以下が井上・柳村（2015）の改訂用の解説文である．

散布図を示す．縦軸は各国の貿易輸出額である．文字どおり桁違いで，下位の国家は数値が重なって国名が読み取れないので，対数を用いて表示した．横軸は外行語使用数で，145 か国について 114 語の Google Trends の数値を合計したものである．縦軸に合わせて対数で表示した．国家分類の記号は以下のとおりである．79 国家にのみ国家名に記号を付けたので，除外した国家・地域は，記号だけで示される．

■主要英語国，●西ヨーロッパ，〇東ヨーロッパ，
▲東アジア，△西アジア，×アフリカ，＋中南米．

記号は，国名ラベルの文字列の先頭に入れた．この文字列をダウンロード済みの「XY Chart Labels」によって，通常のエクセル散布図に貼り付けた．マーカーを消し，ラベル位置を右にしたので，記号の位置のやや左が，プロットされた国家の位置になる．

図 1 の記号を手がかりに大勢を把握する．上と右には■主要英語国，●西ヨーロッパ，▲東アジアが集中する．ここでは外行語も貿易輸出額も多い．中ほどに〇東ヨーロッパ，△西アジア，＋中南米が散在する．外行語も貿易輸出額も中位である．左下には記号だけの国が固まる．大部分が×アフリカで，他に〇東ヨーロッパ，△西アジアの国が混じる．外行語も貿易輸出額も少ない国である．

3.4　近似曲線の入れ方

この散布図に近似直線を入れてみよう．グラフをクリックすると，上端のメニューバーの右側に「レイアウト」が表れる．「レイアウト」をクリックし，「近似曲線」の下の矢印をクリックし，「線形近似曲線」をクリックすると，線が表示される．斜めの線になり，日本語からの外行語の数と貿易輸出額とが相関関係を示すことが読み取れる．

3.5　数式の入れ方

さらに「レイアウト」「近似曲線」の枠の中の「その他の近似曲線のオプション」をクリックすると，「近似曲線の書式設定」が可能になり，下の方に「グラフに数式を表示する」のオプションが表れる．その四角にチェックを入れると数式が出る．「y=2E+08x +

7　「外来語」と逆に外国語に借用された日本語の単語．

8E+09」である．なおこの数式の位置や大きさは自由に変更できる．

この数式は，エクセルの機能上，0 が続くような係数を含むと , 指数形式の誤解しやすい表示になることがある．以前に査読者が「間違っている」と指示をした例である．

《近似直線は y=200000000x + 8000000000 である．エクセル上では係数は各々2E+08, 8E+09 と指数表記されて表示されるが，指数表記の + が加算の + と紛らわしい．なお，指数表記の数値はエクセルのセルに入力し，表示形式を「標準」にすると通常の表記に変換できる.》（井上・柳村 2015 の査読者の補足による.）

3.6　グラフ化の工夫

以上では紙という 2 次元空間を生かした 2 変数組み合わせの散布図を示した．3 変数の合計が 100 % になる場合は，組み合わせて三角グラフ triangram を作成することもできる．第 1，2，3 次産業の構成が典型だが，語種も三角グラフで表せる（和語・漢語・外来語; 混種語はいずれかにまとめる）．言語景観研究に使って，趨勢を論じた例もある（江 2011）.

3 変数の場合は 3 次元グラフとして表示できる．エクセルや SPSS にはないので，STATISTICA などを使うことになる（井上 2004）．これらの視覚化により，新たな流れが読み取れる．なおエクセルの「挿入」「縦棒」「3-D 縦棒」は別の意味で，奥行きの模様がついているにすぎない．また「3-D 縦棒グラフ」は 2 つの項目と 1 つの数値変数を 3 次元に配置できるが，項目を数値変数に変えることはできない.

3.7　直前のデータ保存とフッター

以上のグラフ化では，2 箇所（以上）に保存し，直前のファイルも別名で保存することが望ましい．視覚化は試行錯誤を繰り返すことが多く，前のバージョンを利用することもありうるからである.

またエクセルのフッター機能を最初のグラフに入れておくと，のちに改訂したグラフにも付いて便利である．印刷しておくと，どのファイルのグラフを最終的に論文に採用したかが分かる．フッターは，エクセルのメニューバー左上「ウィンドウズマーク」をクリックし，「印刷」「印刷プレビュー」をクリックし，「ページ設定」の「ヘッダー / フッター」の「フッターの編集」で指定できる．「&[パス]&[ファイル名]&[シート名]&[日付]」を入れておくと，後日グラフを探すのに便利である，

4．河西データの地図対応視覚化の効果

4.1　LAJ 河西データの標準語形使用率

以上は数値のグラフ化についてだった．以下では地図によるデータの視覚化を論じる.

「河西データ」は各都道府県の 82 語の標準語形使用率を示す数値行列である．これまで散布図を作って多様な考察を施した．のちに執筆した鉄道距離についての論文では，都道府県庁所在地間の鉄道距離を用いてグラフ化したことにより，新たな読み取りができた[8].

8　鉄道は過去の街道を踏襲することが多いので，歴史的な交通事情を後世まで反映する．道路距離と違って，カーブを直線化したり新幹線などが通ったりしても，距離を変えないので，追試などに便利である（乗客は運賃を余分に払うが，違法という判決は出なかった）.

後掲図 8 の真ん中の図 b に相当する．当時は散布図に手書きで県庁所在地をつなぐ主要鉄道路線を書き入れた．その読み取り結果をまとめると以下のようだった．

横軸には京都からの鉄道距離，縦軸には標準語使用率を示した．本土は，東京中心の山に見えるが，詳しくみると，二つの山が観察される．一つは京都中心の山で，四国と中国がまとまって，九州にかけてなだらかな山を形成する．日本海縦貫線の傾斜は山陽線から九州にかけてとそっくりである．もう一つ，東海道本線と東北本線は派生した山のように見える．日本海側と太平洋側は離れる．京都中心の方言周圏論的分布と，東京からの近代共通語化の山である．二つの標準語発信地が歴史上存在したことが視覚化されたわけで，この散布図から日本語の「二都物語」を語れる．

4.2　ラベル付加プログラムの効果

散布図で，線や点の数がそれほど多くない場合は，ラベルを付ける手間もかからず，手で入れるので十分だった．しかしデータ数が多いと面倒である．

鉄道距離を用いた以上のグラフができたのは，都道府県名を入れるためのプログラム「XY Chart Labels」が利用できるようになったおかげである．ここでは 3 文字の県略称を用いた（英字 2 字でも 47 都道府県は判別可）．漢字でもいいが，海外での発表も考えてアルファベットを用いた．

4.3　データ反転の地理的効果

ここではもう一つ別の視覚化効果について述べる．図 7 は，縦軸の反転という単純なアイデアの適用結果である．上が方言使用の多い県，下が標準語使用の多い県になる．東日本では，日本海側と太平洋側が離れた．その結果，各県の配置が日本地図と一致するように見える．つまり京都からの鉄道距離が標準語形使用率を支配する．日本海側が西日本の傾斜と線対称をなし，東京関東付近が太平洋にせり出したように見えた．

4.4　散布図のラベル付けと多様な加工手順

図 7 を作るための手順を説明する．図 8 では左端に数表があり，京都からの鉄道距離と LAJ 標準語形使用率と都道府県名が入力してある．右には出力された散布図が 3 枚並んでいる．上から図 a, 図 b, 図 c とする．

数表の B 列，C 列を囲んで，メニューバーの「挿入」「散布図」をクリックする．図 a のような散布図が出る．この段階で，3.8 で前述のフッターを付けるとよい．

数表 B 列，C 列で上下につながるセルの数値は，線付き散布図で線でつながるが，離れたセルの数値はつながらない．北から南まで主要本線を連続したセルにした．数表上部の IBA，GUM，YMG，CIB や，MIE，HOK などは，他の都道府県と主要な鉄道本線でつながっていないと考えて，独立させた．

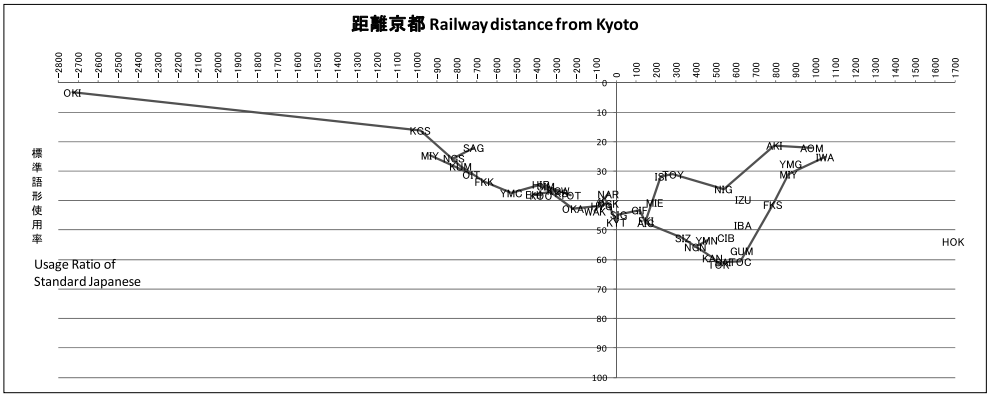

図 7　LAJ 河西データと鉄道距離

線でつないだこのままの図では，都道府県の位置しか示されないので，読み取りが難しい．XY Chart Labels を利用した．図 a の画面をクリックして，メニューバーの「XY Chart Labels」をクリックし，左端の「Add Labels」をクリックすると，「Select a Label Range」の欄が空白として現れる．数表 D 列の 2 行から 64 行（散布図に使った数値部分に対応する都道府県名）を囲むと空白に数式が入る．下端の「OK」をクリックし，しばらく待つと図 a の点にラベルが付く．しかしラベルの文字が大きく，読み取りにくい．

図 a の画面をクリックして，「ホーム」「コピー」「貼り付け」をクリックすると，同じ画面がコピペされる．図 b の位置に置き，画面をクリックして，空色になった枠を上下と左右に拡大すると，文字やマーカーの大きさはそのままで，全体が大きくなる．図 b がその結果である．沖縄と北海道を除くと，京都中心の低い山と東京中心の高い山が重なったように見える．この視覚化により，過去の標準語の普及・伝播の過程に関して，前述の仮説をたてることができた．

図 b をさらに加工すると，別のアイデアに結び付く．図 c は，図 b を「コピー」「貼り付け」した上で，以下の手順を踏んだ．グラフ中央の 0 から 70 の数値の入った軸をクリックし，下端の「軸の書式設定」をクリックし，「軸のオプション」の「軸を反転する」の四角にチェックを入れる．これで上下が逆になった．標準語形不使用率または方言形使用率のグラフである．これは図 b を反転させただけのグラフだが，図 c ではもっと見やすくするために，さらに加工した．まずグラフ右端の「標準語形使用率」の文字をクリックして「削除」した．これでグラフの横幅が広くなった．さらに都道府県名のラベル位置を変えた．任意のラベルをクリックすると，全ラベルが四角で囲まれ，指示の画面が出る．下端の「データラベルの書式設定」をクリックし，「ラベルの位置」を「中央」にした．

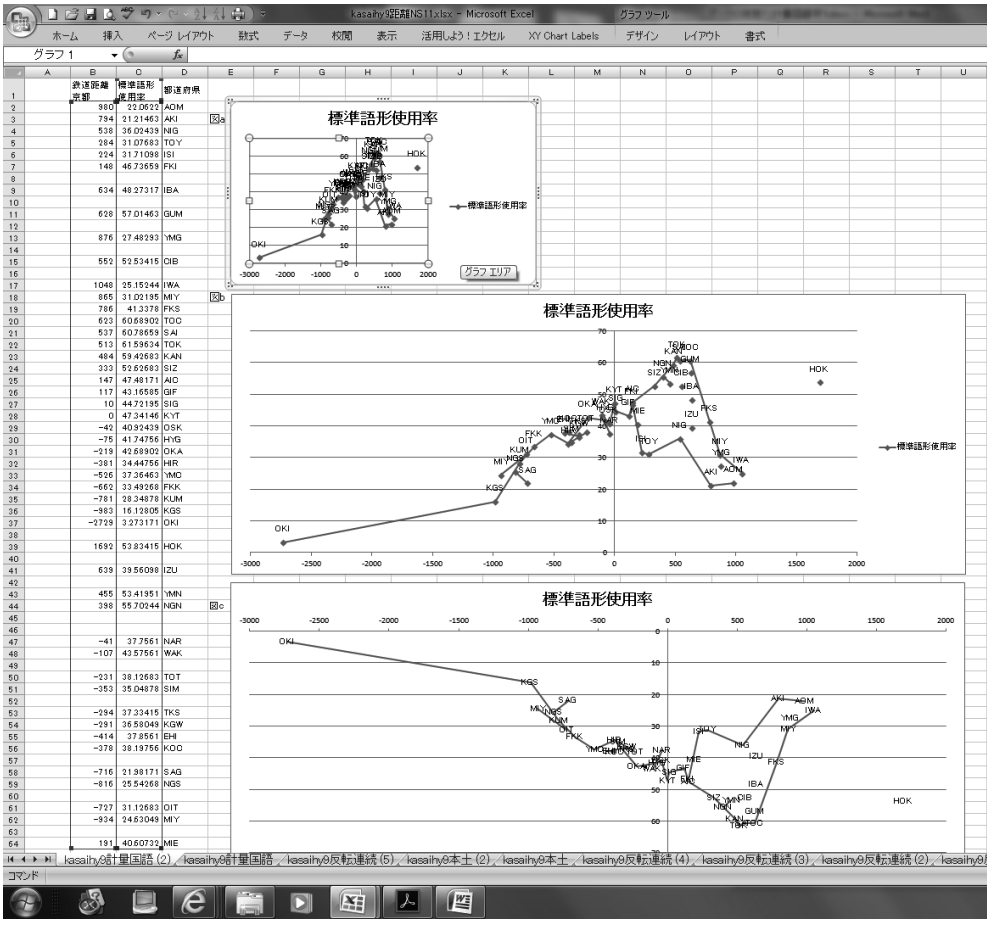

図 8　LAJ 河西データと鉄道距離の作成画面

また，線かマーカーのどこかをクリックすると，指示の画面が出る．下端の「データ系列の書式設定」をクリックし，左側の「マーカーのオプション」を「なし」にした．これで都道府県名ラベルが線でつながる形になり，ラベルと点が離れることがなくなった．最終的に図 7 のようなバイリンガル表示で軸の説明を付けるには，さらに手順が必要だが，林（2015）などでもふれてあるので，省略する．

この手続きによって，都道府県の配置が日本地図と一致して見えたが，この図 7 の視覚化を出発点にすると，さらにアイデアが浮かぶ．太平洋にせり出した関東地方（または東京をピークとする山）を適切に説明するためには，鉄道距離以外が有効である可能性がある．例えば到達時間または時間距離．さらに交通量（トラヒック）全体をとらえるために，特急・新幹線の本数も考慮に入れれば，まさに地理学における重力モデルの適用になる（Chambers and Trudgill 1980, 井上 2001）．ここでは 82 語全体の平均値を用いたが，個々の単語の標準語形使用率に戻って，多変量解析法を適用すれば，鉄道距離や近代の鉄道輸送量などと，関連の深い（または浅い）語の判別もできる．視覚化により，一つの解明が

済むと，さらに分析の視点が広がるという例である．河西データについては，これまで 30 年以上にわたって多様な分析を行った．ここでは計量国語学の趣旨に従って数値のグラフ化について紹介したが，地図化によっても新たな読み取りができた．鑓水（2007）の作成した 20 段階で連続的に使用率を示す日本地図が，考察に役立った．視覚化の効果である．

５．グロットグラム・方言地図への応用

5.1　庄内グロットグラムの新方言

エクセル散布図を使うと，方言地図を描ける．つまり方言地図も（グロットグラムも）広義の散布図の一つなのである．《調査地点位置が散布図グラフとして固定され，そこに様々な項目の回答（語形）が記号として表示される》と位置づけられる．方言地図作成における GLAPS の貢献は大きかった（荻野 1977，1978）．今は作図がもっと楽になった．ことにグロットグラム（＝地理×年齢図）は，地点間の距離と年齢層を等間隔にするなら，エクセルで簡単に作れる．エクセル散布図をもっと活用するなら，地点間の距離を地図上で計算し，かつ生年を軸の値に使えば，地理と年齢を忠実に反映した図を簡単に作れる．

図 9 に庄内グロットグラムの新方言をカウントした合計値を，道路距離に沿って示した．

上下の軸が話者の生年，左右の軸が庄内地方南端を出発点にした道路距離である．話者ごとの新方言の回答を記号（の濃さ）で示した．上端の道路距離 34 キロ付近の，鶴岡市旧市内の若い世代の新方言使用率が高いことが，明瞭に示された．中年以上の使用率も高く，新方言の発生地が地方中都市であることが示された（井上 2014b）．

図中の線は，鶴岡旧市内の新方言の使用率が富士山のように高いことをプレゼンで示すために，挿入した．手順としてはエクセルのメニューバーで「挿入」をクリックし，「図形」をクリックし，「基本図形」の中の「弧」の形を選んで貼りつけ，大きさを調整した．

二重の方言周圏論，つまり京都中心の全国的周圏分布と，その下位に属する鶴岡中心の地方的周圏分布が，新方言（と共通語）で実証された．日本の大部分を覆うような広域グロットグラムでも提示可能である（井上 2015）．

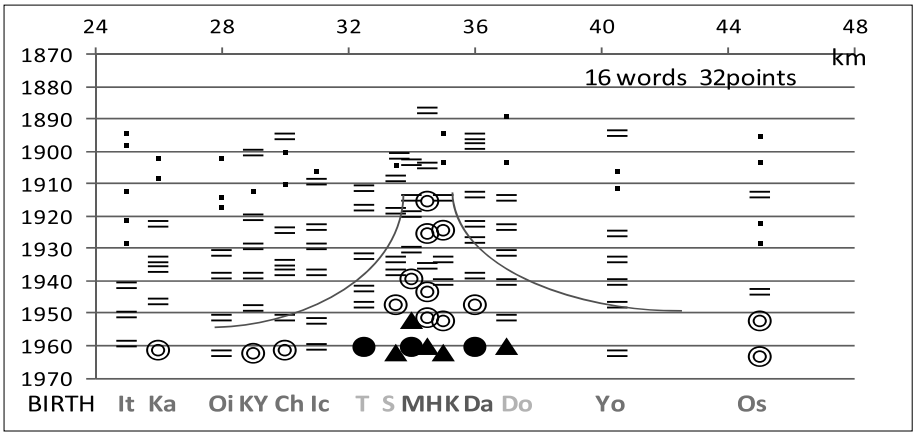

図 9　庄内グロットグラムの新方言と道路距離

5.2　エクセル散布図で作る方言地図

この技法を拡大して，方言地図もエクセル散布図で作れる．まず白地図作成のために調査地点位置を緯度経度で（または LAJ 地点番号を用いて一桁おきに）決定する．（速成としては，単純に地形図での位置の縦と横のセンチを，定規で測ってもいい．）各地点の当該項目の回答を，地点位置の隣の行に入力する．前述の「XY Chart Labels」を使えば，地図上に回答を表示できる．文字で示すと，ジリエロンの ALF（フランス言語地図）と同じで，全体像が読み取りにくい．記号で表したければ，回答をコピーした上で，■●○▲△×＋などの適切な記号に置き換えて，ふたたび「XY Chart Labels」で貼りつければいい．回答を示すラベルにあたるものを記号に置き換えると，自動的にグロットグラムの文字列が記号に置き換わる．全回答に一斉に記号を与えるのでなく，主なものから記号に置き換えるという試行錯誤の手法が，エクセルの「XY Chart Labels」では，適用可能である．

5.3　ステレオ（3D）グロットグラムの可能性

表示を工夫すれば，複数のグロットグラムをまとめて，3 次元で立体的に示すこともできる．グロットグラム調査地域の地図に，地点ごとの柱を立てて，その柱の生年実年代のところに語形をマークすれば，3D の地図になる．小学校の工作の宿題みたいに手作りもできるが，データを入れて，プログラムを利用できれば，3D プリンターで出力できる．パソコン上で動画として見せることも可能になる．また複数の写真を撮って並べれば，ステレオ写真として学術誌に掲載できる．

グロットグラムの多くは出版の年が違うので，話者の生年実年代によって配列すると，語形の広がり方を連続的に把握できる．都染（2012）の多数地域での調査成果のように，男女両方調べてあるときには，実年代が違うだろうから，ますます変化の連続的表示がきれいになる．例えばコーヘンなどをはじめとする新方言の発生と拡大を，立体としてとらえることができる．実行したらおそらく世界最初の試みになるだろう．ただしデータ入力の手間が大変で，白地図に相当する地点と年齢，またある語形を使うかどうかの情報が必要である．このアイデアは庄内地方のグロットグラムのときに適用したが（井上 2003），調査地点がまばらで面白味がなかった．甲南大学の多数のグロットグラム（都染 2012）が一番効果的だろう．実行は難しいが，生データの公開によって，だれかが手がけるかもしれない．

6．まとめ

以上エクセル散布図を利用し，かつラベルを適切に付けることにより，データの内部構造がよりよく分かる例を掲げた．データの視覚化の学術的効果である．ここで掲げたデータは既発表のものである．ちょっとした工夫を加えることにより，考察がさらに深まり，一般理論にまで発展する可能性が秘められている．

散布図の技法を知ったのは，偶然である．他の技法についても，若い研究者に一つずつ教わって覚えた．他に適用できるデータがないかを考えて，少しずつ適用・応用の範囲を広げた．講習や授業を受けたわけでもなく，マニュアルを読んだわけでもないので，体系的に身につけたのではない．ひとりよがりの技法で，もっとエレガントな技法もあるだろ

う．ページ数の制約もあり，手順を画面で示すこともしなかった．メニューバーに出てくることばを「　」で囲んで示した．これまでの体験だと，マニュアルに書いてある日本語はそのままでは理解できず，実際に試して成功した後，はじめて意味が通じることが多かった．自分がそのような文章を書く立場になるとは，思わなかった．不親切な解説であることをお詫び申し上げる．なおここで用いたエクセルのバージョンは Microsoft Office Excel 2007 である．本稿で扱ったグラフの大部分がこれ以前のバージョンによるので，もっと新しいバージョンは試していない．

文献

Chambers, J. K., and P. Trudgill (1980) *Dialectology* Cambridge University Press

井上史雄（2000）『東北方言の変遷』秋山書店 .

井上史雄（2001）『計量的方言区画』明治書院 .

井上史雄（2003）『日本語は年速 1 キロで動く』講談社現代新書 .

井上史雄（2004）「鉄道距離重心と初出年―鉄道距離・使用率・初出年の３Ｄ散布図―」日本語科学 16, pp.47-68.

井上史雄（2011）『経済言語学論考―言語・方言・敬語の値打ち―』明治書院 .

井上史雄（2014a）「ポライトネスの歴史地理学―Google Ngram Viewer と Google trends による言語史―」明海日本語 19, pp.1-10.（web でも公開）

井上史雄（2014b）「昭和の方言―鶴岡と郊外の言語変化―」日本語学 32-15, pp.16-24.

井上史雄（2015）「「お父さん」の記憶時間―グロットグラムによる地域差と年齢差―」社会言語科学 18-1, pp.1-19.

井上史雄・松田謙次郎・金順任（2012）「岡崎 100 年間の「ていただく」増加傾向―受恵表現にみる敬語の民主化―」国立国語研究所論集 4, pp.1-24.（web でも公開）

井上史雄・柳村裕（2015）「外行語世界分布の国別因子分析― Google Trends による傾向―」計量国語学 30-2, pp.73-97.

荻野綱男（1977）「一般的言語地図作成システム GLAPS」計量国語学 11-1, pp19-29.

荻野綱男（1978）「コンピュータ言語地理学 : 言語分析の新手法」言語研究 74, pp.83-96.（web でも公開）

荻野綱男（1988）「日本語における外来語の流入時期と原語」計量国語学 16-4, pp.165-174.

河西秀早子（1981）「標準語形の全国的分布」言語生活 354, pp.52-55.

江源（2011）「言語景観に関する計量的研究」明海日本語 16, pp.71-80.

都染直也（2012）「グロットグラムでみる方言の動態」日本語学 31(13), pp.66-75.

西尾純二（2015）「岡崎市の変容と配慮言語行動」大規模経年調査資料集 21, 国立国語研究所 .（web でも公開）

林直樹（2015）「データの視覚化（2）: Excel によるグラフ作成の基本 (2)」計量国語学 30-2 pp.104-121.

鑓水兼貴（2007）「活用形における共通語の分布パターン―『方言文法全国地図』第 2・3 集データの多変量解析―」計量国語学 26-1 pp.1-18.

横山詔一・真田治子（2007）「多変量 S 字カーブによる言語変化の解析―仮想方言データのシミュレーション―」計量国語学 26-3 pp.79-93.

米田正人（1997）「鶴岡市における共通語化の調査―約 20 年間隔で行われた 3 回の調査を比較して―」日本語科学 2, pp.24-39.

（2015 年 12 月 2 日受付）

Mathematical Linguistics, Vol.30 No.4 (March 2016) pp.216-233.

Tutorial

Data Visualization (4): Application of Excel Scattergram to Graphs and Maps

INOUE Fumio (National Institute for Japanese Language and Linguistics)

Abstract:
Several techniques are discussed here with the idea that linguistic study can be promoted by visualization. Numerical data can be made visible by charts and maps. By this technique, the structure of the data is made easier to read, and new ideas are born. The inner structure of the data can be made clear by applying multivariate analysis technique, and once the structure becomes clear, simpler calculations can be used to indicate the inner structure. There are studies which aim to verify theories and hypotheses, but studies which make the obtained data talk for itself are also necessary. For this, both analysis techniques and representation (visualization) techniques are effective.

Keywords: Excel, scattergram, labeling, XY Chart Labels Tutorial