Vol.2015-CH-107 No.9
2015/8/9

フリーソフトウェア「KH Coder」による計量テキスト分析: 手軽なマウス操作による分析からプラグイン作成まで

樋口 耕一[†1]

概要：KH Coder とは、計量テキスト分析（テキストマイニング）のためのフリーソフトウェアである。内部では茶筌・MeCab・Stanford POS Tagger のほか、MySQL や R を利用しており、これらのツールの機能を統合するために Perl を使用している。本チュートリアルセッションでは第一に、計量テキスト分析の考え方、すなわち KH Coder のフィロソフィーを紹介する。具体的な分析事例を通じて、社会学の分野で伝統的に利用されてきた内容分析（content analysis）の考え方にもとづいた分析方法とソフトウェアであることを示す。第二に、非常に手軽なマウス操作によってテキスト型データの分析が行えることに加えて、Perl ないし R の短いコードを追加すれば新たな分析機能を追加したり、分析を自動化できることを紹介する。これらの点については、ご自身の PC 上で実際に操作を行っていただく予定である。

1. はじめに

本チュートリアルで取り上げる KH Coder とは、テキスト型データを計量的に分析するために筆者が開発・公開したフリー（自由）ソフトウェアである。現在は日本語・英語データを分析できるほか、やや実験的な段階ではあるがフランス語・ドイツ語・イタリア語・ポルトガル語・スペイン語データの分析にも対応している。加えて、中国語データへの対応も現在進めており、年内には中国語データを分析できるアルファ版を公開できる見込みである。操作画面の言語としては、日本語・英語・スペイン語のいずれかを選択できる。2001 年 10 月に最初の版を公開してから改良を続けており、本ソフトウェアを用いた研究事例は、筆者の把握している限りで、学会発表と論文等をあわせて現在 900 件を数えている。

この KH Coder について、本チュートリアルでは第一に、どのような分析を行うために開発したのかという考え方を紹介する。開発の目的や考え方を知っていれば、ソフトウェアの各機能の詳細や使い方をよりスムーズに理解できるだろう。第二に、実際に参加者各自の PC 上で分析を体験していただきながら、KH Coder の使用法を紹介する。ここでは単に既製のソフトウェアを使うことで簡単に分析ができるというだけでなく、KH Coder では様々なカスタマイズが可能なことを強調したい。

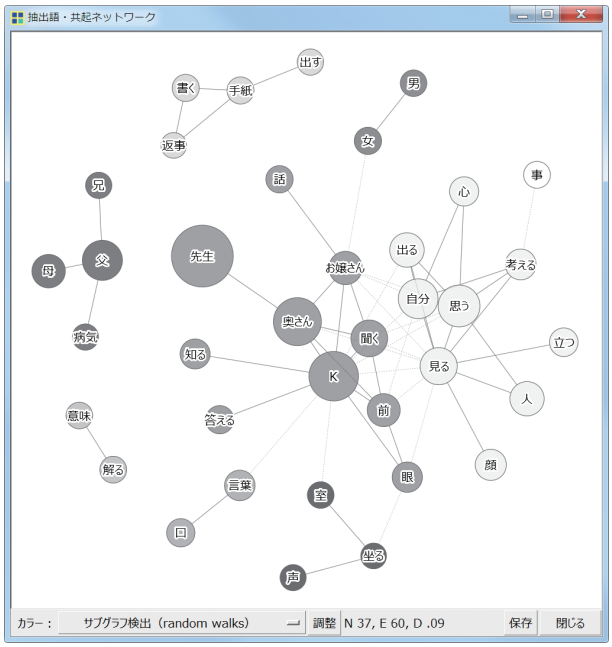

図 1 KH Coder で作成した共起ネットワーク
Fig. 1 Co-occurrence network on KH Coder

2. 計量テキスト分析とは

「文章の微妙なニュアンスを無視して数え上げるなどという分析法は乱暴ではないか」「人間が一字一句の意味をじっくり考えてこそ『深い』洞察を導けるのであり、計量的な分析では『浅い』結果しか得られないのではないか」。人文科学の分野では、テキスト型データを計量的に分析し

[†1] 現在，立命館大学
Presently with Ritsumeikan University

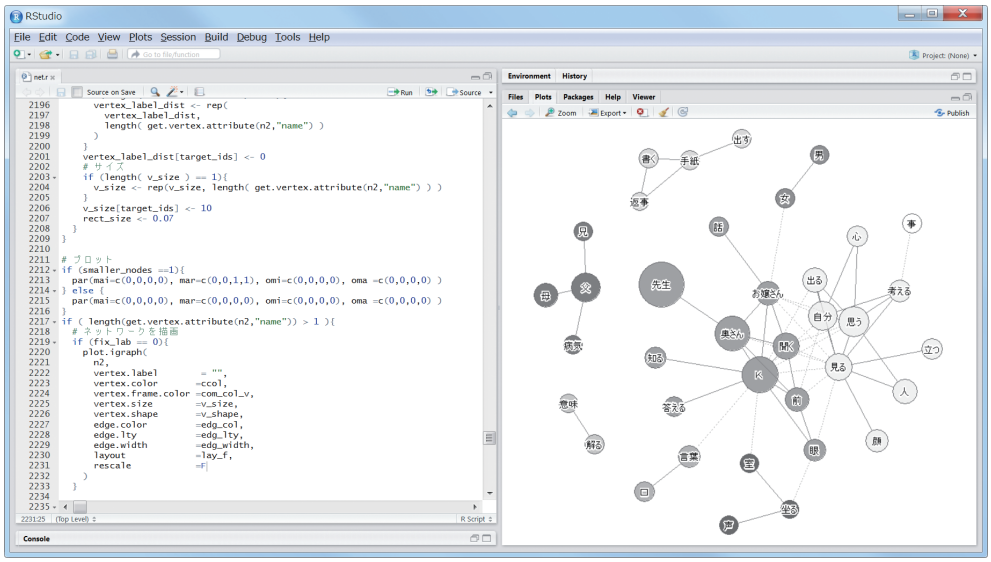

図 2 R コマンドとして保存した共起ネットワークを編集（Rstudio）
Fig. 2 Editing the co-occurrence network saved as R commands (Rstudio).

ようとすると、こうした疑念を突きつけられることがあるかもしれない。仮に計量的な分析を行なうにしても、文章に含まれる唯一の「真の意味」を取り出すことを目指すのか、それとも分析者の観点に応じた多様な解釈を認めるのかを決めなくてはならない。そして、仮に分析者の観点を活かすならば、信頼性ないし客観性をどのように担保するのか。

こうした問題について、内容分析（content analysis）の分野では半世紀以上にわたって議論が蓄積されてきた。そこで、内容分析の考え方に依拠しつつ、近年の自然言語処理・情報技術を活用することで、上述のような疑念ないし問題に答えようというのが計量テキスト分析である。分析者がよりよくデータを理解することを助け、同時に分析の信頼性を向上させることを目指してこの方法を提案している [1]。

3. KH Coder を使った分析

KH Coder ではごく平易な操作で、たとえば図 1 に示すような共起ネットワークを作成できる。さらにこうした分析には、統計解析とグラフィックの環境「R」を用いており、作図のための R コマンドを出力することができる。出力したコマンドをそのまま R で実行すれば、まったく同じ作図を行えるし、R コマンドを編集することで、統計や作図の手法を自在にカスタマイズできる（図 2）。

また、ごくわずかな行数でなおかつ定型的な Perl のプログラムを作成することで、独自の分析コマンドを KH Coder のメニューに追加できる（図 3）。これをプラグインと呼んでおり、プラグインでも当然、R を使った独自の分析や、MySQL を使った独自の検索を行なうことができる [2]。そして、独自の機能を持つプラグインを公開・配布

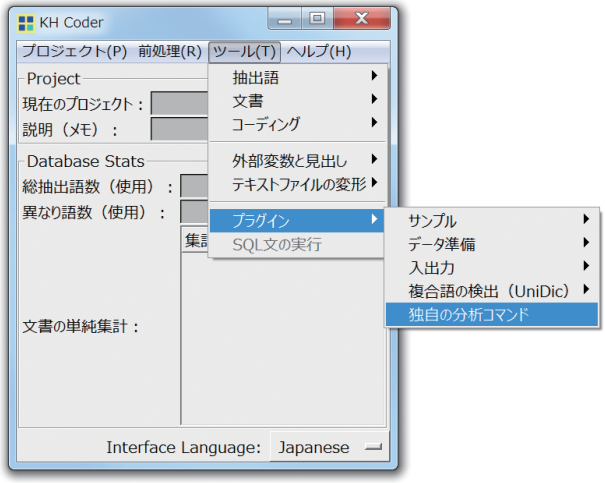

図 3 独自の分析コマンドの追加
Fig. 3 Adding your own analysis function.

することでもできるので、提案手法を普及させたいといった場合にも役立つだろう。

参考文献

[1] 樋口耕一：社会調査のための計量テキスト分析——内容分析の継承と発展を目指して，ナカニシヤ出版 (2014).
[2] 石田基広，神田善伸，樋口耕一，永井達大，鈴木了太：R のパッケージおよびツールの作成と応用，共立出版 (2014).
[3] 阪口祐介，樋口耕一：震災後の高校生を脱原発へと向かわせるもの——自由回答データの計量テキスト分析から，リスク社会を生きる若者たち——高校生の意識調査から（友枝敏雄，編），大阪大学出版会，pp. 186–203 (2015).
[4] 樋口耕一：社会調査における計量テキスト分析の手順と実際——アンケートの自由回答を中心に，コーパスとテキストマイニング（石田基広，金　明哲，編），共立出版，pp. 119–128 (2012).
[5] 石川慎一郎，前田忠彦，山崎　誠（編）：言語研究のための統計入門，くろしお出版 (2010).