2019 年 10 月 16 日

「データの読み込み」など補足 3 題

KH Coder の新しいバージョンに関して本書（「やってみようテキストマイニング」）と関連する内容をはじめ、以下の 3 つのテーマについて補足的な解説を行います。

1 ：データの読み込み方法

2 ：Excel などへのエクスポート機能

3 ：仮説コーディングによる分類問題

1．データの読み込み方法（本書 3 章）

本書で解説している自由回答アンケートのデータは，基本的にテキスト部と外部変数で構成されます．これらを KH Coder に読み込む標準的な方法は，まとめて一括して読み込むやり方です．一方本書（の 3 章）では 2 つの部分を分割して読み込む方法をとっています．

最初に Excel ファイルから一括して読み込むやり方を本書の事例である「高齢者向けサービス」のデータを使って説明します．

（1）データの一括読み込み

読み取るデータが Excel 上で図 1 のように入力されているとします．このファイルは本書のサイト上からダウンロードできます．（注）KH Coder では Excel の最初の sheet から読込まれます．

	A	B	C	D	E	F	G
1	No	性年代	性別	就業形態	子供の有無	家族構成	高齢者向けサービス
8	7	男性55-59歳	男	3	2	1	時間はあると思うので社会貢献できるサークルがあり、地域貢献できる環境があれば良いと思います。
9	8	男性55-59歳	男	1	2	3	具体的なアイディアは浮かばないが、基本的には、機械やテクノロジーに頼ることなく、心の触れ合いを重視したサービスがあれば良いと思う。
10	9	男性55-59歳	男	1	1	4	混雑時の高齢者専用電車
11	10	男性55-59歳	男	1	2	2	宗教を介さずに死ぬことに対する恐怖心を軽減してくれるサービス
12	11	男性55-59歳	男	2	1	5	高齢者向けの医療サービスやレクレーション等が充実した施設
13	12	男性55-59歳	男	3	2	2	日常生活の支援
14	13	男性55-59歳	男	3	2	1	年金生活者への公的な機関の生活資金の貸しだし
15	14	男性55-59歳	男	5	1	3	終活
16	15	男性55-59歳	男	3	1	3	若い人との交流
17	16	男性55-59歳	男	3	1	4	同行支援
18	17	男性55-59歳	男	3	1	4	家事や買い物の代行など、日々生活で必要なサービスが無料もしくはわずかな費用で受けられるようになること。
19	18	男性55-59歳	男	3	2	1	買い物の足
20	19	男性55-59歳	男	4	1	5	定期的に様子を見に来てくれるようなこと
							高齢者をただ見守って介護するだけでなく、リハビリ・ス

図 1.「もとデータ」の一部

このファイルを KH Coder の新規プロジェクトとして設定し，図 2 のようにして「分析対象とする列」でテキスト部の項目を指定します．図 1 のデータの場合には G 列の「高齢者向けサービス」が分析対象になります．この項目以外の A 列「No」から F 列「家族構成」までは外部変数として扱われます．

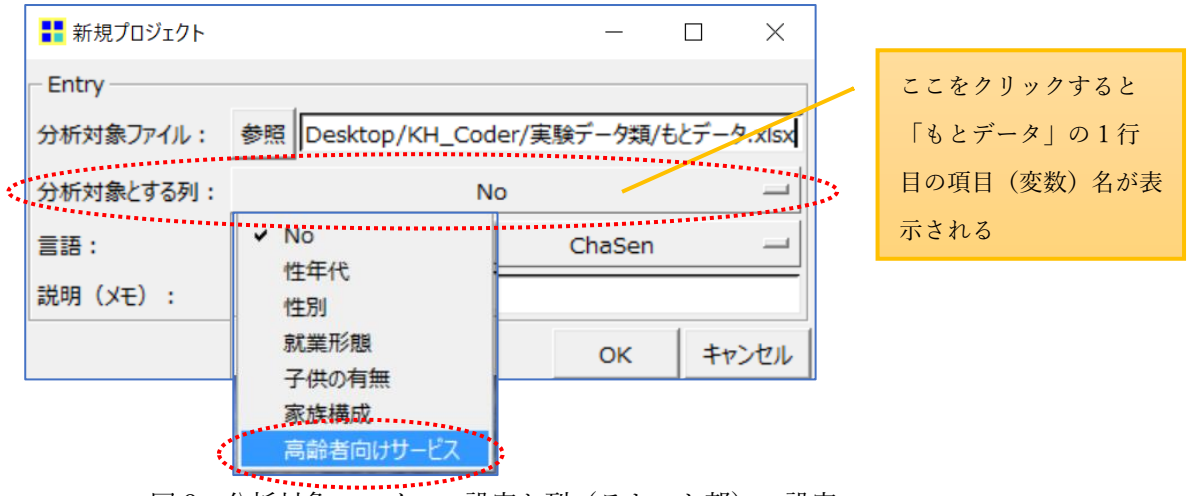

図 2．分析対象ファイルの設定と列（テキスト部）の設定

KH Coder にデータが取込まれると，読み込んだ Excel ファイルと同じフォルダ内に「***_txt0」と「***_var0」という名称の 2 つのテキストファイルが作成されます．2 つのファイル名最後の“0”はバージョンを表しています．これらのファイルの内容は以下のようです．

①テキスト部（FA 部）のテキストファイル（***_txt0）

「分析対象とする列」として指定したテキスト部のファイルは，図 3 のようなファイルとして作成されます．1 行ごとに<h5>---cell---</h5>という行が挿入されます．KH Coder の各種の検索や分析の最小単位は H5 になります．本書の 2 章 2.4 節で解説しているような，Excel ファイルの 1 つのセル内に改行コードがあったとしても問題は起こりません．

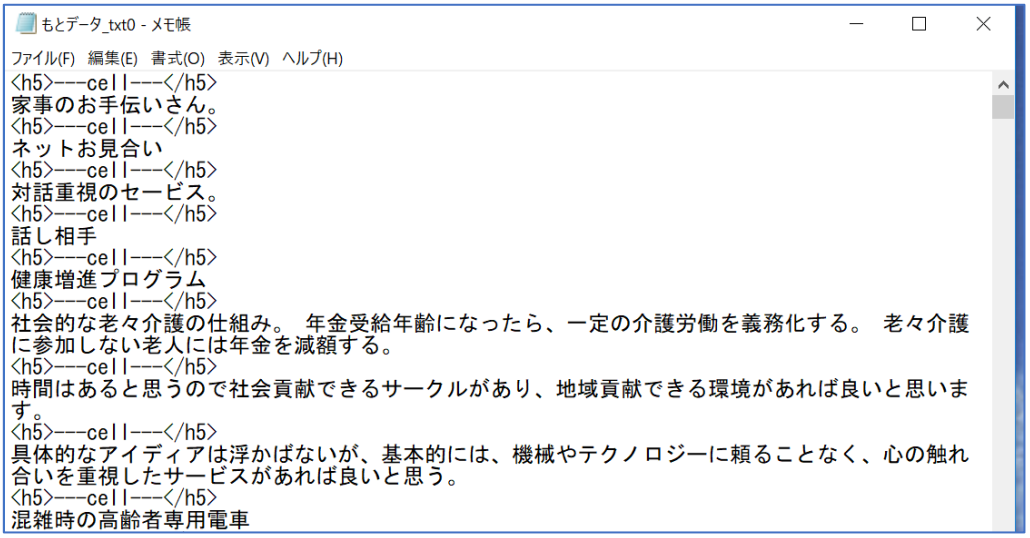

図 3.テキスト部のファイル

②外部変数のテキストファイル（***_var0）

一方，テキスト部以外の項目は図 4 のようなファイルとして作成されます．①のテキスト部のファイルと異なり，外部変数名（“No”～“家族構成”）が 1 行目に取込まれます．本書のようにテキスト部を読み込んだ後に，改めてメインメニュー「ツール／外部変数と見出し」から外部変数を読み込む必要はありません．

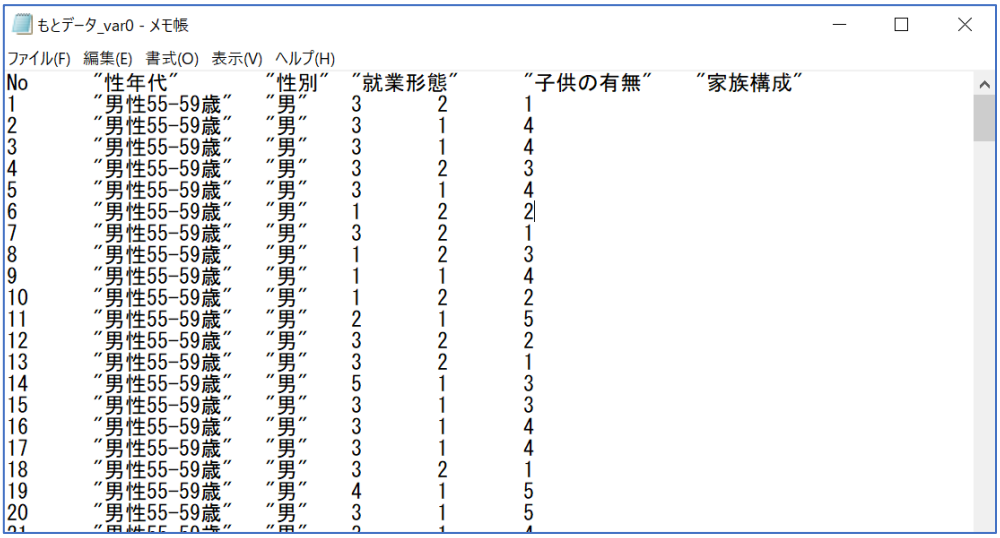

図 4. 外部変数のファイル

（2）本書の分割読み込み方法

本書は，アンケート調査の自由記述の部分をテキストマイニングすることを前提として解説している本です．アンケート調査の FA 部を分析する場合には，本書の第 2 章や 3 章で説明しているとおり，時間が許す限り，何度も編集作業を繰り返しながらデータの精度を高めていく必要があります．一方，外部変数に関しては有効サンプルを抽出し，カテゴリーの合成などの基本的な編集をした後ではほとんど修正作業は必要ありません．

さらに，編集した FA 部のテキストファイルを Excel 等の他のソフトで直接利用することもしばしばあります．次項 2 で説明する Excel などへのエクスポートファイルを利用する場合などが典型的な利用シーンです．

こういったことが 2 種類のファイルを別々にして読み込んでいる理由です．

2．Excel などへのエクスポート機能（本書 3.4，4.1.1，B.1，6.4）

KH Coder には検索や分析結果を Excel ファイルなどにエクスポートする機能がありますが，新しいバージョンではメインメニュー「プロジェクト」のサブメニューのひとつに集約されました（図 5）．

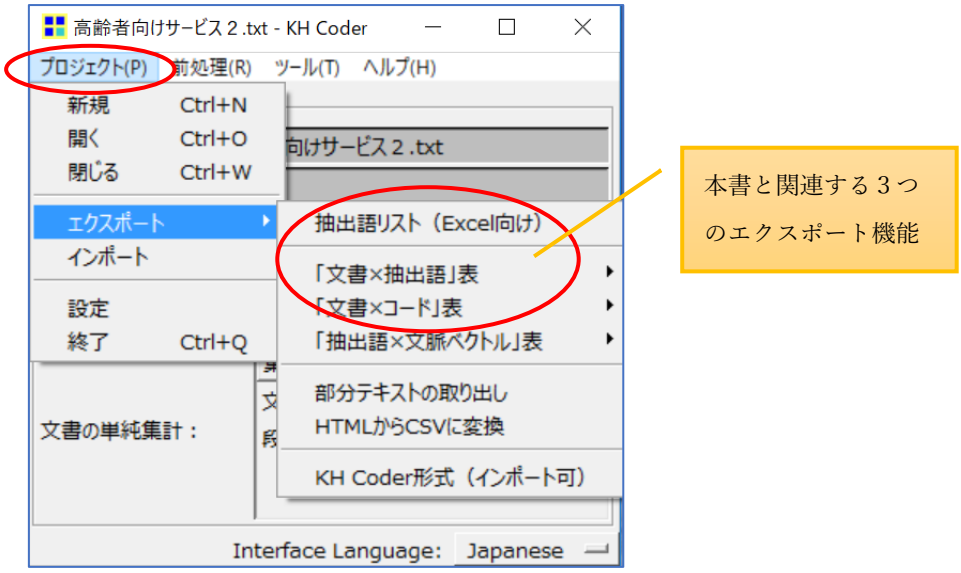

図 5．集約されたエクスポート機能

本書と関連する部分は，図 5 に示した以下の 3 カ所です．

①抽出語リスト（Excel 向け）

「3.4 抽出語の暫定的なリスト表示」と「4.1.1 抽出語リスト－抽出語を頻度順に並べる」で解説している部分です．

②「文書×抽出語」表

付録「B.1 クロス集計用 Excel データ」において，抽出語を Excel ファイルに出力し

て利用している部分です．本書 145 ページの図 B.3 はこの機能を利用して出力した Excel ファイルです．

③「文書×コード」表

「6.4 コーディング結果の出力と利用法」において，本書 111 ページの図 6.21 に示している表はこの機能を使って出力したものです．

３．仮説コーディングによる分類問題（本書第 6 章）について

本書第 6 章では KH Coder の第 2 段階の分析について解説しています．いろいろな分析が可能と思われますが，本書では仮説コーディングの機能を利用して，いくつかのテーマを定義・設定し，文書全体を分類し理解するための方法を紹介しています．

仮説コーディングの機能を活用することの有効性を次頁の図 6 に整理しましたが，その前に，コーディングルール・ファイルを利用することに関して，KH Coder の開発者である樋口耕一先生は，その著書（の後半マニュアル部）で，以下のように説明されています．

KH Coder マニュアルの一部

（略）なお，コードを付与する条件の指定内容によっては，1 つの文書が複数の条件に該当するということが起こりうる．この場合には，1 つの文書に対して複数のコードが付与される．というのもKH Coder によるコーディングは，「犯罪」か「合法」かのどちらか一方といった，排他的なカテゴリーに文書を分類するという処理ではない．むしろ，文書の中から要素を取り出すという考え方の処理である．1 つの文書がたとえば「犯罪」と「人情」のような複数の要素を含むことはありうるという前提にもとづいている．（略）

クラスター分析やベイズ学習による分類法は，ひとつのサンプルを単一のカテゴリー（クラスター）に分類しようとする方法ですが，コーディングルール・ファイル利用する場合には，複数のコードに対応づけることができることになります．本書で解説しているアンケート調査などの場合には，そもそも 1 人の自由回答をひとつのカテゴリーだけに無理矢理分類できないということが少なくありません．その意味で仮説コーディングによる分類は非常に柔軟性があります．というよりも現実的な多重の分類が可能であると言えます．

また，サンプルによっては，分析者が定義したカテゴリーに必ずしも分類できない場合もあります．それを確認することは，分析者が気づかなかった未定義の新しい仮説あるいはテーマの発見につながるかもしれません．これは自由回答を分析することの最も大きな効用のひとつと言えます．

さらに，コーディングルール・ファイルは，ベイズ学習による分類と同様の使い方ができる点も重要です．過去に実施した調査データの分析用に作成したコーディングルール・ファイルは，例えば定期的に実施されるその後の調査データの分析にすぐさま利用することができます．そこで未定義のテーマが発見できれば，それまでの傾向との違いとして認識でき

ることになるでしょう．時系列的な傾向の分析が可能になります．

そしてさらに，分析者によって定義された新コードは，文書全体を要約し理解することを目的とした分析のための新しい軸であり変数になりますが，ある意味では多変量データの次元縮小を目的として適用される主成分分析における主成分や因子分析における因子と同じ役割を果たします．しかしながら，それらの手法との大きな違いは，変数間の相関情報によって機械的に抽出されたものではないという点です．テキストマイニングにおける新コードは，分析者が解釈し，仮説としての意味を持つ変数として構成されたものです．分析者にとっては分析者自身の思いが反映できる非常に強力なサポート機能であると思われます．是非，第 2 段階の分析まで進めてみましょう．

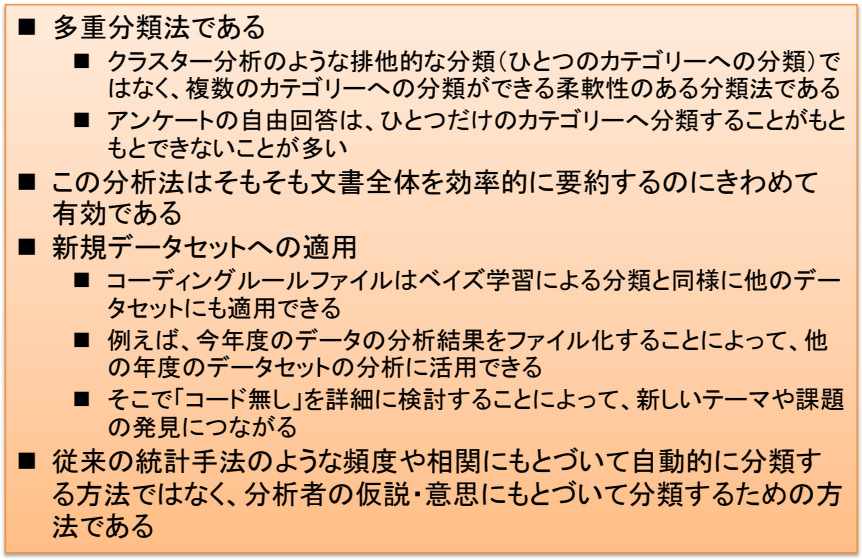

図 6．仮説コードの有効性