

# A Two-Step Approach to Quantitative Content Analysis : KH Coder Tutorial using *Anne of Green Gables* (Part I)

HIGUCHI Koichi<sup>i</sup>

**Abstract :** This article introduces a two-step approach to performing quantitative content analysis of text data. First, an outline of the approach is briefly described. Second, the procedure of using the approach to analyze the novel *Anne of Green Gables* is described as a tutorial. Third, the features of the approach are discussed with reference to the results of the analysis.

The tutorial section of this article allows readers to simulate the same analysis on their own personal computers. We use free software and most of the necessary operations are illustrated in figures. The subject of the analysis is the popular novel *Anne of Green Gables*. It is pointed out that the heroine Anne's foster mother Marilla plays an essential role in the novel and that Marilla is more important than Anne's best friend Diana, and Gilbert with whom Anne has a faint romance. In the analysis of the tutorial, we confirm whether the quantitative analysis based on the two-step approach also illustrates the importance of Marilla.

The first half of this article is published here. It is planned that the second half will be published in this bulletin in the near future.

**Keywords :** quantitative content analysis, KH Coder, Anne of Green Gables, tutorial

## 1 Introduction

### 1.1 Two-Step Approach

This article introduces a two-step approach to quantitative content analysis of text data. Content analysis has been extensively employed to analyze qualitative data, such as text in the field of social sciences and humanities. In this article, first, an outline of the two-step approach is described. Second, the procedure of applying the approach to the novel *Anne of Green Gables* is described as a tutorial, allowing readers to simulate the same analysis on their own personal computers. Third, the features of the approach are discussed with reference to the results of the analysis.

The author has proposed a quantitative content analysis approach that comprises the following two steps (Higuchi 2004, 2014).

- Step 1: Extract words automatically from data and statistically analyze them to obtain a whole picture and explore the features of the data while avoiding the prejudices of the researcher.
- Step 2: Specify coding rules, such as "if there is a particular expression, we regard it as an appearance of the concept A", and extract concepts from the data. Then, statistically analyze the concepts to deepen the analysis.

---

<sup>i</sup> Associate Professor, Faculty of Social Sciences, Ritsumeikan University

The analysis procedure of Step 1 is similar to the method “text mining” and is performed almost automatically. The same results can therefore be obtained no matter who analyzes the data, and the results are hardly contaminated by the prejudices or hypotheses of the researcher. Meanwhile, it is sometimes difficult to use one’s own perspective or pursue one’s own research questions. In such cases, the researcher can proceed to Step 2 and perform coding, which is a procedure conventionally used for content analysis. By performing coding, the researcher can take a closer look at any aspect of interest in the data.

## 1.2 KH Coder: Practical Free Software

To allow anyone to easily carry out analysis by adopting the above two-step approach, the author has been developing and distributing free software called KH Coder. The software could analyze Japanese text only when it was first published in 2001. Currently, in addition to English and Japanese, it supports Catalan, Chinese, French, German, Italian, Korean, Portuguese, Russian, Slovene, and Spanish text. As far as the author knows, more than 1000 studies using KH Coder have been published as of November 2016. While most of these studies have been published in Japanese, more than 100 studies have been published in English.

KH Coder uses Stanford POS Tagger to extract words from English data, R for statistical analysis, and MySQL to organize and retrieve the data. These software programs, including KH Coder, have been used by many researchers. Additionally, since the source code is open to the public<sup>1</sup>, anyone can check what the software does if necessary. In other words, KH Coder is not a closed black box but is open to verification by third parties. This openness is desirable especially for academic uses.

## 2 *Anne of Green Gables* as the Subject of Analysis

### 2.1 Purpose of Analysis

The novel *Anne of Green Gables* describes how an orphan, Anne, grows up after being adopted by a family comprising 60-year-old Matthew and his younger sister Marilla. Anne becomes good friends with Diana, a girl of the same age living in the neighborhood, and competes with a boy named Gilbert at school. Anne speaks and laughs a lot and is gradually accepted as a member of the family.

It has been pointed out that the foster mother Marilla plays an important role in this story.

The development of the story really follows the education of Marilla. The relationship between Anne and Marilla is the central, most complex relationship in the novel, to which even Anne’s relationship with Matthew and Diana (and with Gilbert, which follows behind all of these) must yield (Doody 1997).

Doody (1997) states that Marilla is more important than Anne’s best friend Diana, and Gilbert with whom Anne has a faint romance. Doody (1997) focuses on the changes in Marilla, who gradually learns to love a child through her experience of bringing up a little girl, Anne, and points out that the changes in Marilla are the center of the story<sup>2</sup>.

The main purpose of conducting the analysis described in this article is to confirm whether the quantitative analysis can also illustrate the importance of Marilla. Additionally, readers who happen to have read *Anne of Green Gables* will see the analysis results of a known story, and so will be able to confirm the features and reliability of the analysis approach by comparing the results with the story they remember. Furthermore, the most efficient way for readers to learn how to analyze data using KH Coder is to simulate the same analysis as described in this article on their own personal computers. It is also useful to check how the results change if the reader chooses options other than the examples given in this article in the analysis window.

## 2.2 Preparation of Data

To analyze text data using KH Coder, you must prepare the data as a plain text file or as an Excel file in the format shown in Figure 1.

	A	B	C
1	text	chapter	part
2	Mrs. Rachel Lynde is Surprised	01	01-07
3	Mrs. Rachel Lynde lived just where the Avonlea main road	01	01-07
4	There are plenty of people in Avonlea and out of it, who	01	01-07
5	She		-07
6	And yet here was Matthew Cuthbert, at half-past three	01	01-07
7	Had it been any other man in Avonlea, Mrs. Rachel, de	01	01-07
8	"I'll just step over to the Gables after tea and find ou	01	01-07
9	Acco		-07
10	"It's just STATING, that's what, she said as she stepped	01	01-07

Figure 1: Preparing data (tutorial\_en\anne.xls)

In the data used here (Figure 1), the text of *Anne of Green Gables* is entered in the first column (column A). Each cell is filled with one paragraph in the same order as the paragraphs are written in the original novel. This column is named “text”. The chapter number that contains the paragraph is entered in each cell in the second column (column B), which is named “chapter”. Moreover, the whole story is divided into four parts (Chapters 1 to 7, 8 to 19, 20 to 28, and 29 to 36), and the part that contains the paragraph is entered in the “part” column (column C).

In KH Coder, columns such as “chapter” and “part” are called “variables”. By preparing these columns, you can find the characteristic words of each chapter or part (Figure 11). Additionally, when you retrieve a sentence, you can check which chapter and which part contain the sentence (Figure 7). Preparing useful information as variables will greatly help your analysis.

## 3 Installation and Setup of KH Coder

### 3.1 Download and Installation

This section describes the installation procedure assuming a personal computer running a Windows operating system; the procedure does not apply to Linux or Macintosh operating systems. Linux and Macintosh users should refer to the relevant part of the manual rather than this section. Once installation is completed, however, the procedure for analysis, which is described from the next section, applies to all Linux, Macintosh and Windows platforms.

It is recommended to install spreadsheet software in advance for browsing tables created by KH Coder. You can use free software such as LibreOffice or OpenOffice as well as Microsoft Excel.

At present, KH Coder can be downloaded from <http://khc.sourceforge.net/en/>. The latest version at the time of writing is 3.Alpha.08. Download the file for Windows (\*.exe file) from this URL and unzip it as

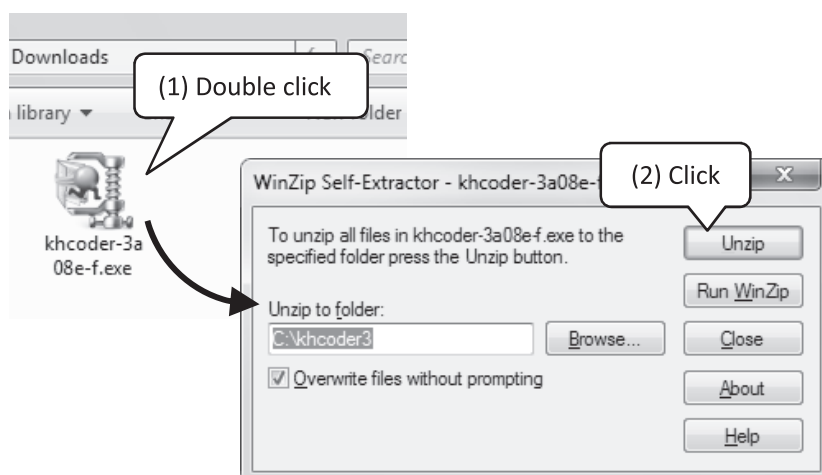


Figure 2: Installing KH Coder

shown in Figure 2. Then, double-click the unzipped “kh\_coder.exe” to start the software (Figure 3)<sup>3</sup>.

If the menu and other text in the KH Coder window are displayed in Japanese, look for the indication of “Interface Language: Japanese” on the lower right of the window. This text is always displayed in English. To change the interface language to English, click “Japanese” in this text and change it to “English”.

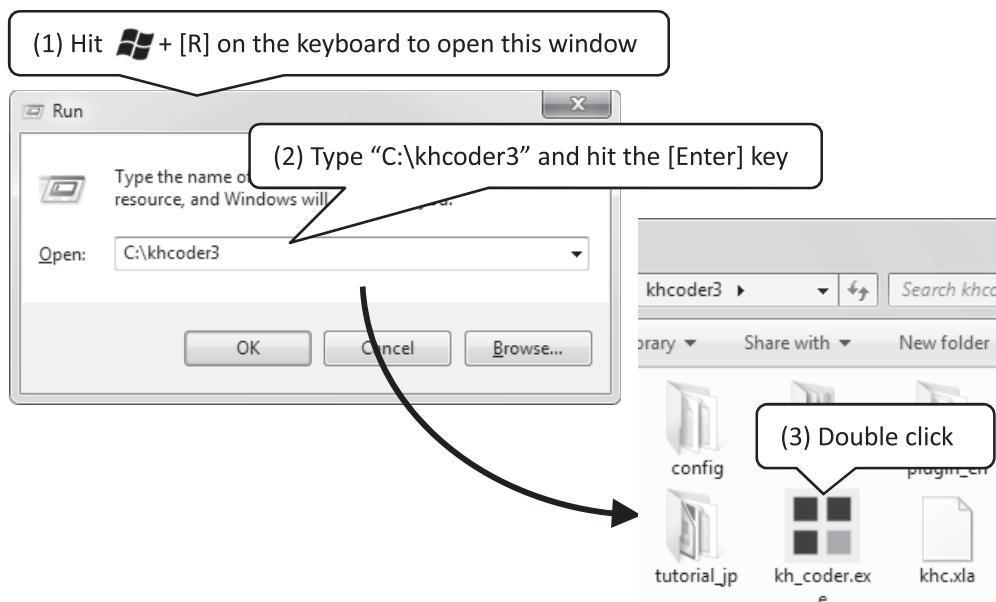


Figure 3: Starting KH Coder

### 3.2 Designating English Stop Words

Common words found in all kinds of writing, such as “a”, “an”, and “the”, are not important words for content analysis. It may be desirable to remove such words from analysis results, such as a word frequency list. KH Coder allows users to remove such words from the scope of analysis and retrieval by designating them as “stop words”. Figure 4 shows the procedure for designating stop words.

In Figure 4, the example list of stop words included with KH Coder is used. However, the words to be designated as stop words may vary depending on the purpose of analysis and data. In such cases, you can add and delete stop words in the window shown in Figure 4<sup>4</sup>.

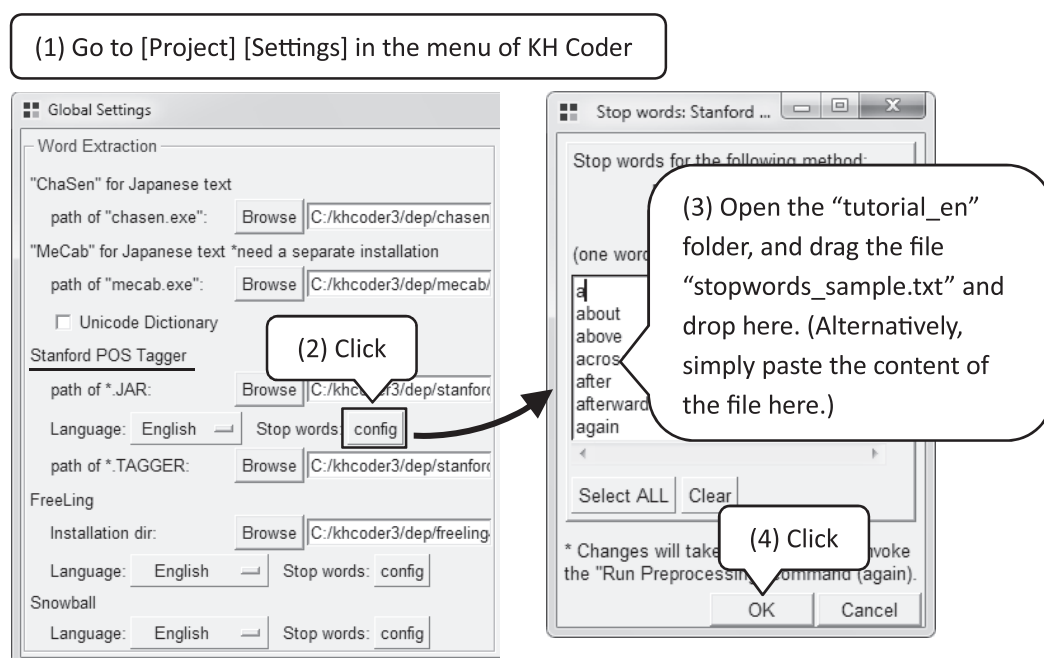


Figure 4: Designating English stop words

## 4 Step 1: Overview of the Novel

### 4.1 Creating a Project and Pre-processing

To perform analysis using KH Coder, you should register the data file to be analyzed in KH Coder as a “project” and execute pre-processing. Be aware that if you move or delete the data file after creating a project, you will be unable to continue analysis.

The project creation procedure from (1) to (4) shown in Figure 5 only needs to be carried out once. Afterward, a list of already-created projects will be displayed if you click “Project”, then “Open” on the menu after starting KH Coder, allowing you to choose the project for which you want to resume analysis from the list.

By procedure (5) shown in Figure 5, words are extracted from the data and processed into a database with their POS names identified. This processing may take several tens of seconds. You can start actual

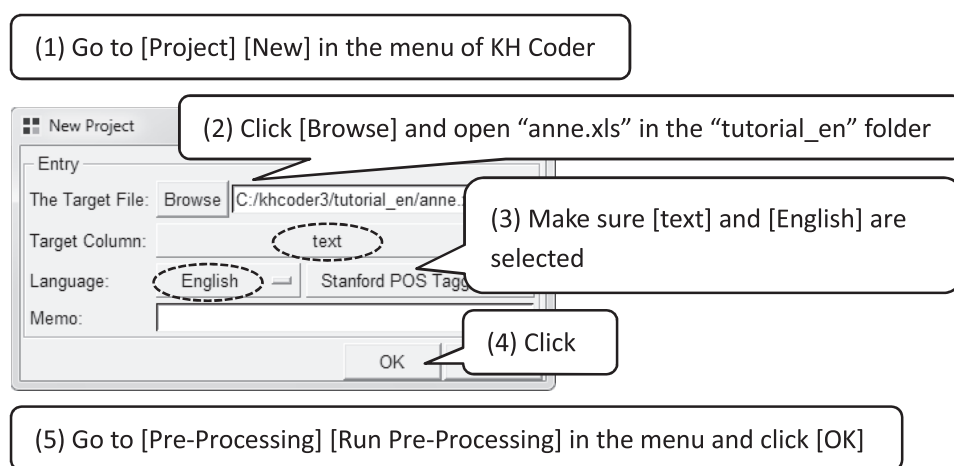


Figure 5: Preparing for data analysis

analyses after the pre-processing is completed.

#### 4.2 Frequent Words and Their Contexts

As the first step of analysis, let us check the words frequently appearing in *Anne of Green Gables*. By the procedure shown in Figure 6, a list of the 150 most frequently occurring words is displayed. Table 1 shows the top 30 words among them; the words to which the author pays most attention are hatched or underlined in this table.

Table 1 shows that the words representing the main characters appear frequently: "ANNE" appears 1138 times, "MARILLA" 849 times, "Diana" 414 times, and "Matthew" 361 times. The character name that most frequently appears next to the heroine "ANNE" is not her best friend of the same age "Diana" but Anne's foster mother "MARILLA". Furthermore, the frequency of "MARILLA" (849 times) is more than double that of "Diana" (414 times). Judging only from the appearance frequency, it is obvious that Marilla plays an important role in this story. Additionally, judging from the fact that the number of occurrences of

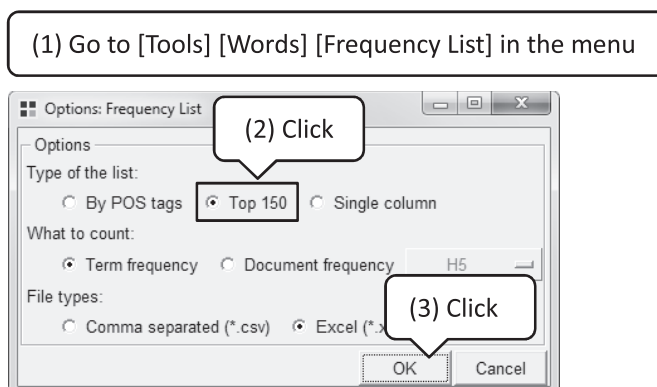


Figure 6: Creating a word frequency list

**Table 1: List of the 30 most frequent words**

Words	Freq	Words	Freq	Words	Freq
ANNE	1138	little	283	want	149
say	952	girl	267	home	136
MARILLA	849	thing	260	child	134
think	486	tell	252	Barry	132
Diana	414	look	246	school	128
know	364	good	225	sit	126
Matthew	361	feel	215	night	117
just	358	time	208	really	116
come	353	eye	152	hair	114
make	286	Lynde	151	Gilbert	113

“Gilbert” is less than that of “Lynde”, which is the family name of Anne’s neighbor, Gilbert is considered to play only a limited role. However, such an issue should be discussed not only on the basis of the number of occurrences but also through a more detailed analysis.

Table 1 also shows several common words such as “say” and “think”, which are likely to appear frequently in any story. Additionally, some other words help us infer the theme of the story. They represent that an orphan “girl” or “child” heroine gets adopted, finds a “home”, and goes to “school”, and represent the color of her “hair” about which she once had a complex.

When interpreting any analysis result of KH Coder, not limited to a word frequency list such as Table 1, be sure to confirm how each word is used in the original data. Even if you obtain analysis results such as “a certain word frequently appears” or “a certain word is characteristic of a certain part”, they mean nothing unless you understand the meaning of the word in the specific data you analyze. This is because even the same word may have different meanings in different contexts or usage. In addition, if the list contains a strange word that makes you think “Why does such a word appear frequently?” or “Why is this word listed as a characteristic word?”, there may be a chance of making a discovery. By investigating how such a word is used in the text, you may discover characteristics of the data that you had not realized before.

Considering the above, the author has developed functionalities not only for statistical analysis but also for flexible data retrieval. Among such retrieval functions, the KWIC concordance is convenient for checking the context in which the word is used. Figure 7 shows how to retrieve data using this function. The “Document” window displaying the whole paragraph (Figure 7) also shows the values of the variables that indicate in which chapter and which part the paragraph is contained.

Note that KH Coder extracts and counts every word after converting it to its original form. For example, the 952 occurrences of “say” in Table 1 include those of “say”, “says”, “saying” and “said”. Furthermore, retrieval is performed for the original forms in principle. This is the reason why the search results of “child” also include those of “children” in Figure 7. Unless the settings are modified, the words designated as stop words, prepositions, conjunctions, and the like are excluded from the scope of analysis and retrieval.

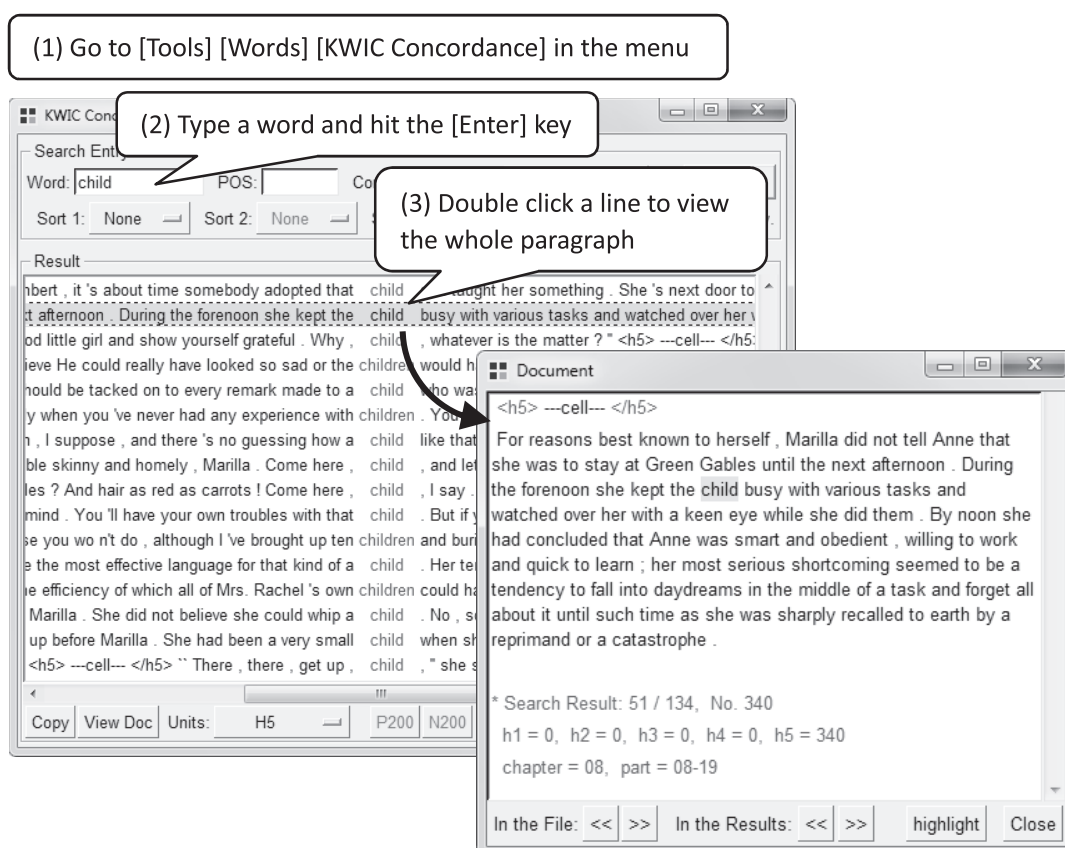


Figure 7: Checking the context where the word is used

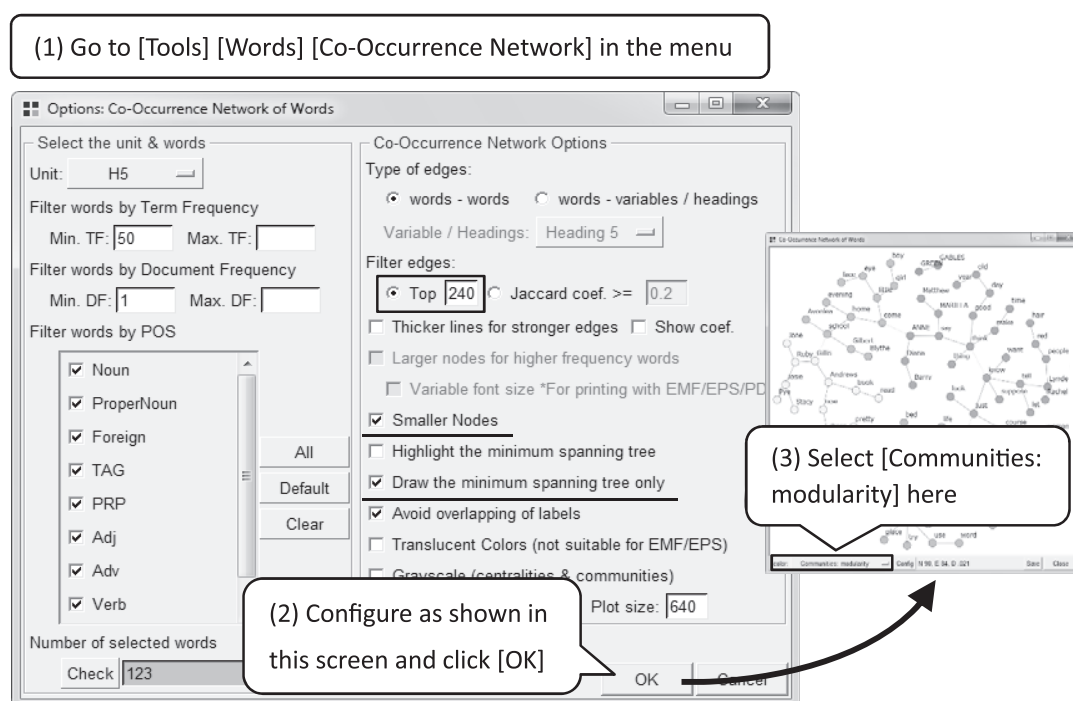
#### 4.3 Co-occurrence Network of Words

We next explore what words are used together frequently by generating a co-occurrence network of major words. Generally speaking, you will be able to read the main themes of the data by seeing the groups of frequently occurring words that are often used together. For example, if the three words “new”, “dress”, and “pretty” frequently co-occur in the data, it can be supposed that there is a theme of fashion or dressing up in the data. In the case of *Anne of Green Gables* data, you can also see the links between the characters to infer the role of each character.

The co-occurrence network has been traditionally used in content analysis to statistically express the data (Osgood 1959, Danowski 1993). In this procedure, we visualize the co-occurrence structure in data by drawing a network connecting words that tend to be used together. Since it is a network, we must see whether words are connected by lines. There is not much meaning to the positions of words. Even if two words are nearby, it does not mean that the degree of co-occurrence is strong unless those words are connected by a line<sup>5</sup>.

Figure 8 shows the procedure for generating a co-occurrence network using KH Coder. In Figure 8, 123 frequently occurring words that appear 50 times or more are designated as the scope of analysis. By default, KH Coder generates a network by connecting 60 pairs of the most strongly co-occurring words by lines. Words not connected by lines are removed from the result diagram. To display more co-occurrences





**Figure 8: Generating a co-occurrence network**

(lines) and words, change the number in the box indicated as “Top 240” in Figure 8. In Figure 8, this number is changed from the default value of 60 to 240. In some cases, however, if you increase the number of co-occurrences (lines) to be displayed as above, the lines may be concentrated on a small part of the network depending on the data structure, making the results hard to read. In such cases, you can make the part with dense lines easier to read by using the “Draw the minimum spanning tree only” option. In the co-occurrence network displayed as a result, several groups of words strongly connected with one another are automatically detected and displayed with different colors. In Figure 9, to make it easy to distinguish the groups in black and white, the boundaries between the groups are indicated with thick dashed lines and each group is given a number in parentheses.

Although we have configured detailed options in Figure 2, it is a good idea to just click the “OK” button without changing the option and look at the result in actual analysis. We may then try increasing the number of co-occurrences (lines) to be displayed and selecting other options to see how the result changes. Without repeating the operation as shown in Figure 8 from the beginning, we can change options by clicking the “Config” button on the screen displaying the result. Through trial and error, it would be beneficial to pursue a visualization suited to the characteristics of the data and the purpose of analysis. A diagram that contains necessary information, that is easy to read, and that has functional beauty would be ideal.

Regarding the links between the characters, “Diana”, “Marilla”, and “Matthew” are connected close to “Anne” in Part (1) of the co-occurrence network (Figure 9). This suggests that the story depicts the close relationships between Anne and her best friend Dianna, foster mother Marilla, and foster father Matthew, whereas “Gilbert” is in rather remote Part (9) and connected to “Anne” via “school”. “Jane”, “Ruby”, and



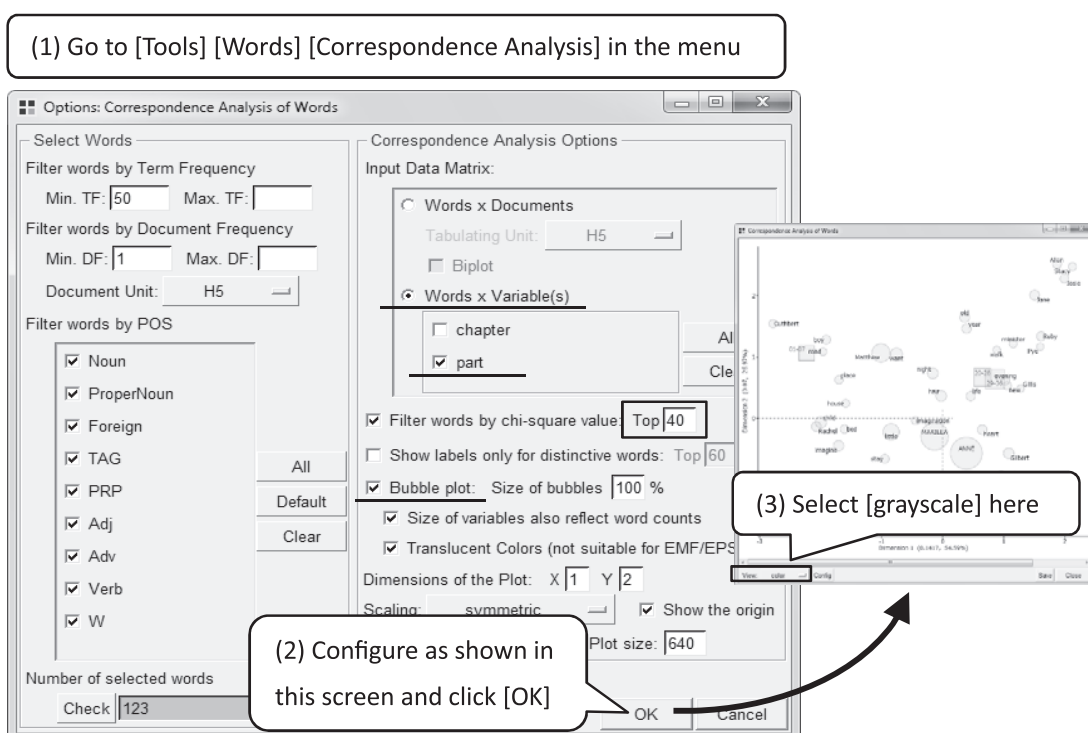


Figure 10: Executing correspondence analysis

data into four parts and visualizes the characteristic words of each part employing correspondence analysis (Greenacre 2007).

Figure 10 shows the procedure for correspondence analysis. The correspondence analysis uses Column C (i.e., the “part” column) in Figure 1. This column includes four kinds of values, such as “01-07” and “08-19”. For example, “01-07” means that the text in that line is contained in one of Chapters 1 through 7. These columns are called variables in KH Coder. In Figure 10, “Words x Variables” and then “part” are selected. The analysis uses 123 frequently occurring words that appear 50 times or more and especially focuses on 40 words for which the number of appearances appreciably changes between parts<sup>6</sup>.

Figure 11 shows the results of correspondence analysis. The values of the “part” variable, such as “01-07”, and words, such as “Anne” and “Marilla”, are plotted with squares and circles, respectively. Using correspondence analysis, you can explore the correspondence between the variable and words by plotting them on the same diagram. The area of each circle is proportional to the number of occurrences of each word. Therefore, the more frequently the word appears, the larger the circle becomes. The area of each square is proportional to the number of words in the text of that value.

In correspondence analysis, uncharacteristic words uniformly found in all parts are plotted near the origin (0, 0) (i.e., the point at which the ordinate and abscissa are both zero) whereas words having strong characteristics are located away from the origin. For example, “Cuthbert” is plotted far from the origin in the upper left of Figure 11, meaning that the word has strong characteristics. We then ask, what are the characteristics of the word “Cuthbert”? The word is far away in the direction of “01-07” as seen from the origin, which means the word appears especially frequently in part “01-07”. Reading the characteristics of

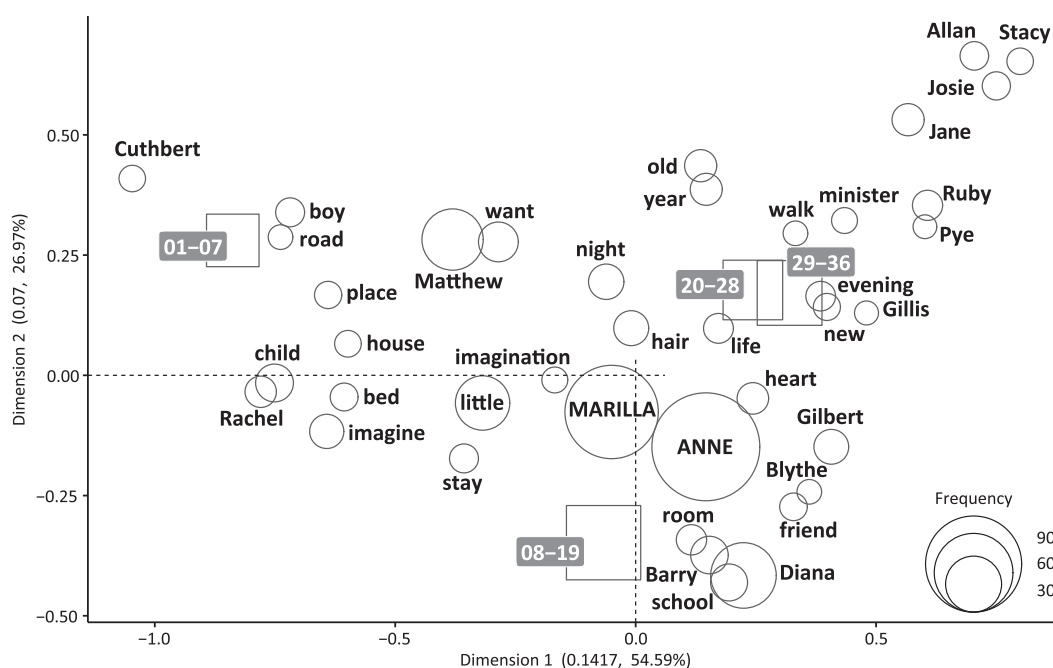


Figure 11: Correspondence analysis of words and variables

each part from the words plotted in a similar direction as seen from the origin as above is the basic way of interpreting correspondence analysis.

In addition, it is also effective to see where each value of the variable is located. In Figure 11, “01-07” and “08-19” are away from other values, but “20-28” and “29-36” are close to each other. This means that frequently occurring words are similar for “20-28” and “29-36”, which suggests that these two parts have similar contents.

We can read the following characteristics of each part by viewing the words plotted in Figure 11 as described above and seeing how they are used in the text. First, part “01-07” describes the livelihood of the “Cuthbert” siblings, Marilla and “Matthew”. They decided to adopt a “boy” from an orphanage to help them run their farm, but a girl named Anne was sent by mistake. At this stage, Anne is often called “child” instead of by her own name. Anne, who likes to “imagine” various things, is eventually allowed to “stay” with the “Cuthbert” family.

Next, in “08-09”, Anne becomes a good friend with “Diana” “Barry”, a girl living in her neighborhood, and starts going to “school”. At “school”, “Gilbert” teases Anne about her red hair, so she comes to hate him. Later, in “20-28” and “29-36”, Anne and Diana go separate ways, and Anne’s schoolmates, such as “Josie”, “Jane”, and “Ruby”, become characteristic. Anne also learns a lot through interactions with adult women such as Mrs. “Allan”, a wife of a minister, and Miss “Stacy”, Anne’s school teacher.

The orphan Anne is accepted first by the “Cuthbert” family, next by the “Barry” family in her neighborhood, and eventually by the local society including her school. Figure 11 shows that the story as a whole progresses in this way. The fact that “MARILLA” is located near the origin in Figure 11 indicates that she appears almost evenly throughout the four parts.

#### 4.5 Closing Remarks for Step 1


In the previous sections, we confirmed frequently appearing characters and words from the frequency list (Table 1) and saw links mainly between the characters in the co-occurrence network (Figure 9). We also read the flow of the story throughout the novel from the correspondence analysis (Figure 11). Marilla appears far more frequently than all other characters except the heroine Anne (Table 1), and her relationship with Anne appears to be almost as strong as Diana's (Figure 9). She appears not sporadically but throughout all four parts of the story (Figure 11).

Even from only the results of Step 1 of the analysis, which automatically extracts words and statistically analyzes them, we can understand the importance of Marilla to some extent. In Step 2, the researcher can focus on an aspect of his/her own interest to see what role Marilla plays in the story in more detail.

The analysis so far has not interpreted the meanings or roles of all the words in the figures and tables. This does not mean that they are omitted because here *Anne of Green Gables* is just an example for the tutorial. Even in actual research, we cannot interpret the meanings of all the words, because the words in the figures and tables are extracted mechanically, and therefore inevitably include words not related to the purpose of the analysis and words that do not interest the researcher. Although it is desirable to interpret as many words as possible, it is impossible to interpret all 50 or 100 words including such irrelevant words.

Consequently, what words draw attention and how they are interpreted will vary largely depending on the researcher's interest, while anyone can obtain the same figures and tables if the settings are the same. Since this is not automatic summarization but analysis, such variation naturally occurs depending on the researcher's point of view. Additionally, such variation will lead to creative and original analysis.

#### Notes

- 1 A source code is a form of software that is easy for a human to check and edit. Source codes are normally kept secret in the case of commercial software.
- 2 The importance of Marilla is discussed also in Japan (Kawabata 2008, Matsumoto 2008, Yamamoto 2008).
- 3 The Windows key () used in the operation shown in Figure 3 (1) is normally located near the lower-left corner of the keyboard. If you cannot find the Windows key, start Explorer and open "C:\khcoder3" manually.
- 4 The designated stop words can be restored to the scope of analysis and retrieval as follows. Click [Pre-Processing] and then [Select Words to Analyze] from the menu of KH Coder, check "OTHER" in the "parts of speech" pane in the window displayed, and then click "OK". The words designated as stop words are given a special POS (part of speech) name of "OTHER". Those words having a POS name of "OTHER" are removed from the scope of analysis and retrieval unless the above operation is performed.  
In addition to the words manually designated as stop words, words not representing the contents of the writing, such as prepositions and conjunctions, are automatically classified as "OTHER". Please refer to the manual for details of the POS system of KH Coder and how to modify the system.
- 5 Positions of words are arranged to make the network easy to see. For example, it is better for lines connecting words to intersect or overlap as little as possible. Because we use random numbers to compute this positioning, the word placement may differ depending on the version of KH Coder and the operating system. However, even though the positioning of words varies, which words are connected by lines or the grouping result shown in Figure 9 does not change. Refer to the manual for details of KH Coder's algorithm for generating co-occurrence networks.
- 6 As with the co-occurrence network above, in actual analysis, it is a good idea to first click on "OK" and see the result without changing detailed options. In this example, once you select the variable "part", you can click "OK" without changing anything else. After that, you can try other settings, like selecting "Bubble plot" and

reducing the number of words from the default “60” to “40”. With such small trial and error, the quality of results can be appreciably improved.

### References

- Danowski, J. A., 1993, “Network Analysis of Message Content”, W. D. Richards Jr. & G. A. Barnett eds., *Progress in Communication Sciences IV*, Norwood, NJ: Ablex, 197-221.
- Doody, M. A. 1997, “Introduction”, W. E. Barry, M. A. Doody & M. E. D. Jones eds. *The Annotated Anne of Green Gables*, Oxford University Press, New York, 9-34.
- Greenacre, M. J., 2007, *Correspondence Analysis in Practice 2nd ed.*, Boca Raton, FL: Chapman & Hall/CRC.
- Higuchi, K., 2004, “Quantitative Analysis of Textual Data: Differentiation and Coordination of Two Approaches”, *Sociological Theory and Methods*, 19(1): 101-15 (Written in Japanese).
- Higuchi, K., 2014, *Quantitative Text Analysis for Social Researchers: A Contribution to Content Analysis*, Nakanishiya Publishing: Kyoto, Japan (Written in Japanese).
- Iker, H. P. & N. I. Harway, 1969, “Computer Systems Approach toward the Recognition and Analysis of Content”, G. A. Gerbner, O. R. Holsti, K. Krippendorff, W. J. Paisly & P. J. Stone eds., *The Analysis of Communication Content: Developments in Scientific Theories and Computer Techniques*, New York: Wiley & Sons, 381-486.
- Kawabata, Y., 2008, “Surprise of Marilla Cuthbert” Katsura, Y. and Shirai, S. eds. *The world of Masterpieces We Want to Know More 10: Anne of Green Gables*, Minerva: Kyoto, Japan, 109-19 (Written in Japanese).
- Matsumoto, Y., 2008, *Journey to the Anne of Green Gables: Hidden Love and Mystery*, NHK Publishing: Tokyo, Japan (Written in Japanese).
- Osgood, C. E., 1959, “The Representational Model and Relevant Research Methods,” I. d. S. Pool ed., *Trends in Content Analysis*, Urbana, IL: University of Illinois Press, 33-88.
- Osgood, C. E., G. J. Suci & P. H. Tennenbaum, 1957, *The Measurement of Meaning*, Urbana, IL: University of Illinois Press.
- Saporta, S. & T. A. Sebeok, 1959, “Linguistic and Content Analysis,” I. d. S. Pool ed., *Trends in Content Analysis*, Urbana, IL: University of Illinois Press, 131-50.
- Stone, P. J., 1997, “Thematic Text Analysis: New Agendas for Analyzing Text Content,” C. W. Roberts ed., *Text Analysis for the Social Sciences*, Mahwah, NJ: Lawrence Erlbaum, 35-54.
- Yamamoto, S. 2008, *From Anne Shirley to Jane Eyre: Introducing English Literature in University Classes*, University of Tokyo Press: Tokyo Japan (Written in Japanese).

## 接合アプローチによる量的内容分析の実践（一） —『赤毛のアン』を用いた KH Coder チュートリアル—

樋口 耕一<sup>i</sup>

本稿では、量的な内容分析を実践するための方法として筆者が提案している「計量テキスト分析」を、新たな分析事例とともに紹介する。計量テキスト分析において、データを分析する具体的な手順にはいくつかのバリエーションがあるが（Higuchi 2014）、本稿では特に「接合アプローチ」と呼ばれる手順を取りあげる。第一に、このアプローチと、その実現のために筆者が開発・公開しているフリーソフトウェア KH Coder について概要を手短に紹介する。第二に、このアプローチにもとづいて小説『赤毛のアン』を分析する手順を、読者が自分の PC で同じ分析を行えるチュートリアルの形で記述する。第三に、分析の結果を踏まえて、本アプローチの特徴について議論する。

本稿で紹介する接合アプローチとは、従来の内容分析で利用されてきた2つのアプローチを接合したものである。従来の内容分析では、テキスト型データを計量的に分析するために Correlational アプローチか Dictionary-based アプローチを用いることが多かった。Correlational アプローチはクラスター分析のような統計手法を用い、頻繁に同じ文書の中にあられる言葉のグループを見つけだすといった方法で、データ中の主題を探索するアプローチである。このアプローチは Statistical Association アプローチと呼ばれることもある。それに対して Dictionary-based アプローチでは、統計手法ではなく、分析者自身の指定した基準にそって言葉や文書を分類し、計量的な分析を行なう。これら2つは考え方が大きく異なるアプローチでありながら、実際の分析においては混同されやすい部分もあった。そこで混同されやすい部分を峻別した上で、これら2つを接合したものが、本稿で紹介する接合アプローチである。

本稿のチュートリアルでは、この接合アプローチを用いて、小説『赤毛のアン』の原文を分析する。小説『赤毛のアン』では、主人公である孤児のアンが、マシューとマリラの兄弟に引き取られ、成長していく様子が描かれている。この物語においては養母マリラの果たした役割が非常に大きいという指摘がある。親友のダイアナや、アンとの淡いロマンスが描かれるギルバートよりも、マリラの方が中心的であったという。また『赤毛のアン』は、マリラが子供を愛することを学び、それによって自分自身も幸せになっていくという、大人の成熟と生き直しの物語であると指摘されている。本稿の分析では、こうしたマリラの重要性を、計量的分析からも読み取ることができるのかどうかを確認する。

なお本稿の前半をここに掲載する。後半については本誌の将来の号に掲載の予定である。

キーワード：量的内容分析, KH Coder, 赤毛のアン, チュートリアル, 計量テキスト分析

---

i 立命館大学産業社会学部准教授