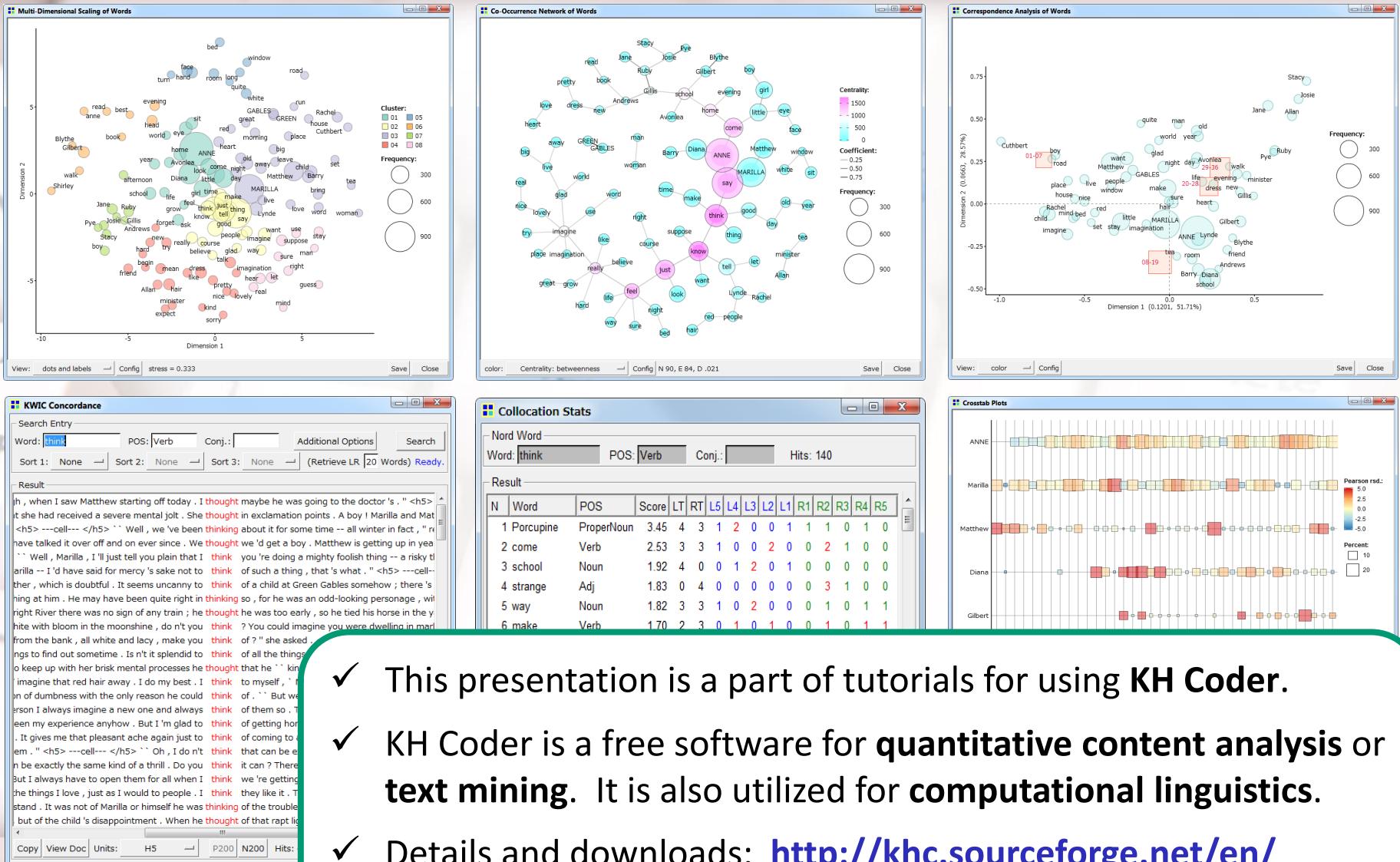


# KH Coder Tutorial using *Anne of Green Gables*:

A Two-Step Approach to Quantitative Content Analysis

Koichi Higuchi

# Preface



- ✓ This presentation is a part of tutorials for using **KH Coder**.
- ✓ KH Coder is a free software for **quantitative content analysis** or **text mining**. It is also utilized for **computational linguistics**.
- ✓ Details and downloads: <http://khc.sourceforge.net/en/>

# Table of Contents

- Introduction
  - ✓ [Data](#)
  - ✓ [Purpose of Analysis](#)
- Preparation
  - ✓ [Install KH Coder](#)
  - ✓ [Configure Stopwords](#)
  - ✓ [Create a Project and Run Pre-Processing](#)
- Step 1
  - ✓ [Word Frequency List](#)
  - ✓ [The Context where a word is used](#)
  - ✓ [Co-occurrence Network of Words](#)
  - ✓ [Correspondence Analysis of Words](#)
  - ✓ [Closing Remarks for Step 1](#)
- Step 2
  - ✓ [Use Coding Rules to Count Concepts](#)
  - ✓ [Retrieve Documents Assigned a Specific Code](#)
  - ✓ [Characters in Each Chapter](#)
  - ✓ [Characters and Verbs](#)
  - ✓ [Change of Words Co-occurring with Marilla](#)
  - ✓ [Conclusions](#)

# Introduction

# Data

- We are going to analyze a novel *Anne of Green Gables* by Montgomery.
- When you prepare your own data for analysis, please open the attached “Anne.xls” file in “tutorial\_en” folder and see the figure below.

	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 ro	01	01-07
4	There are plenty of people in Avonlea and out of it, who	01	01-07
5	She was surprised to find that there were so many people in Avonlea -07		
6	And yet there was Matthew Cuthbert, at half-past three, who	01	01-07
7	Had it been any other man in Avonlea, Mrs. Rachel, de-07		
8	"I'll just step over to Mrs. Barry's," said Anne. "I'll have a cup of tea and find out about the new school." -07		
9	According to Anne, the new school was to be opened in Avonlea. (* ) Enter data in the first sheet if you use Excel or Calc -07		
10	"It's just starting, that's what," she said as she stepped into the room. -07		

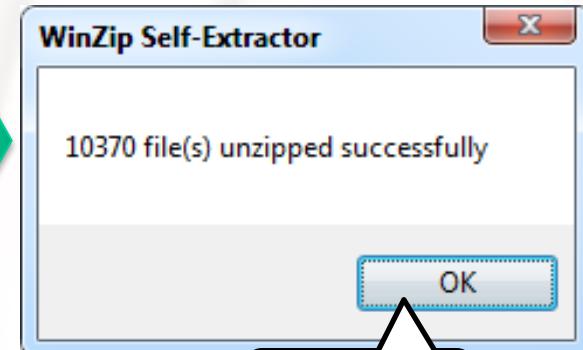
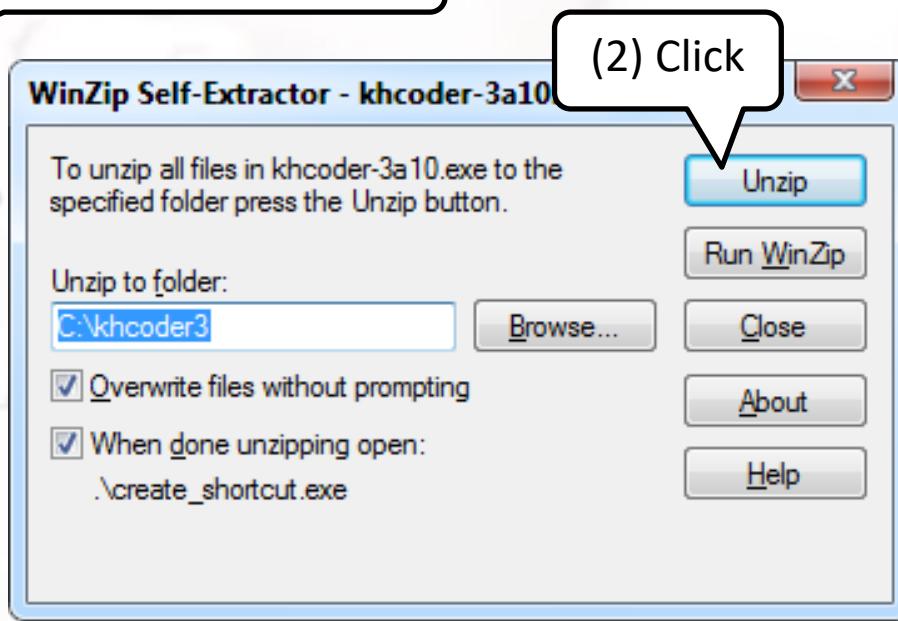
# Purpose of Analysis

- To confirm whether the quantitative analysis can also illustrate the centrality of Marilla
  - ✓ It has been pointed out that the heroine Anne's foster mother Marilla plays an essential role in the novel and that Marilla is more central than Anne's best friend Diana, and Gilbert with whom Anne has a faint romance.
- To demonstrate a quantitative content analysis approach that comprises the following two steps:
  - ✓ [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.

# Preparation

# Install KH Coder

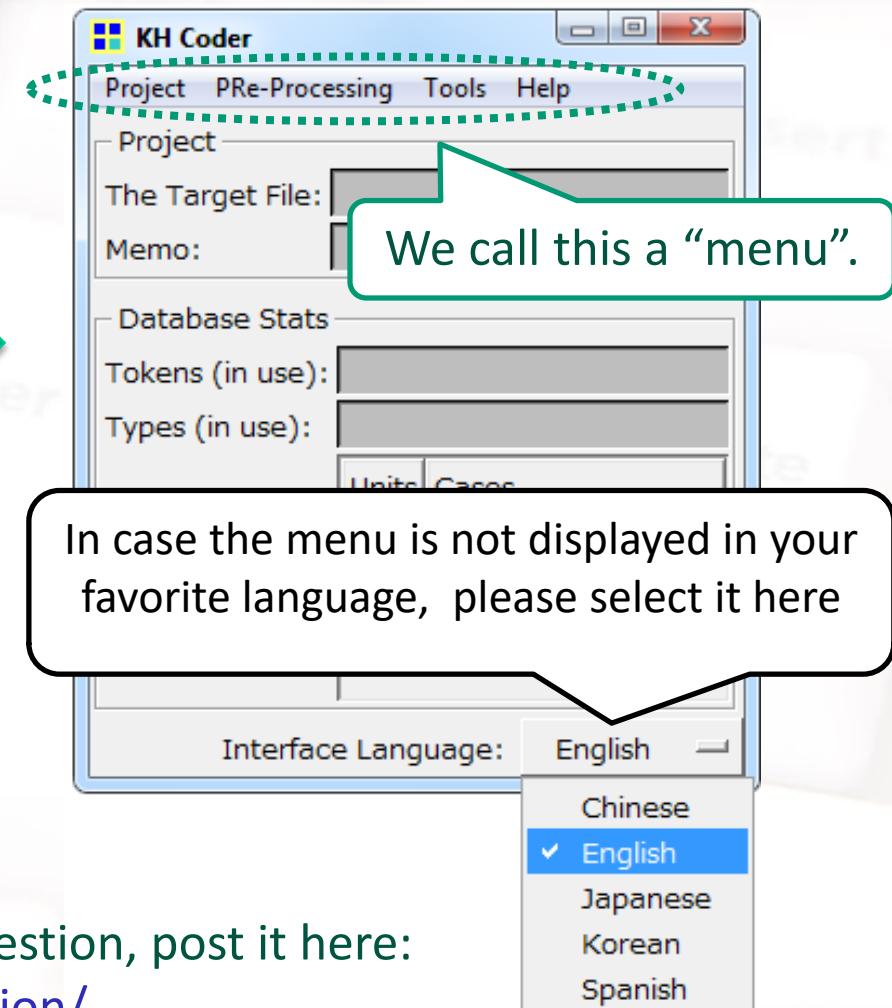
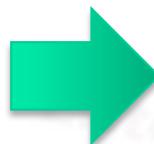
(1) Double click the downloaded file



- Now you are ready.
- The number of unzipped files may vary between versions.

# Interface Language

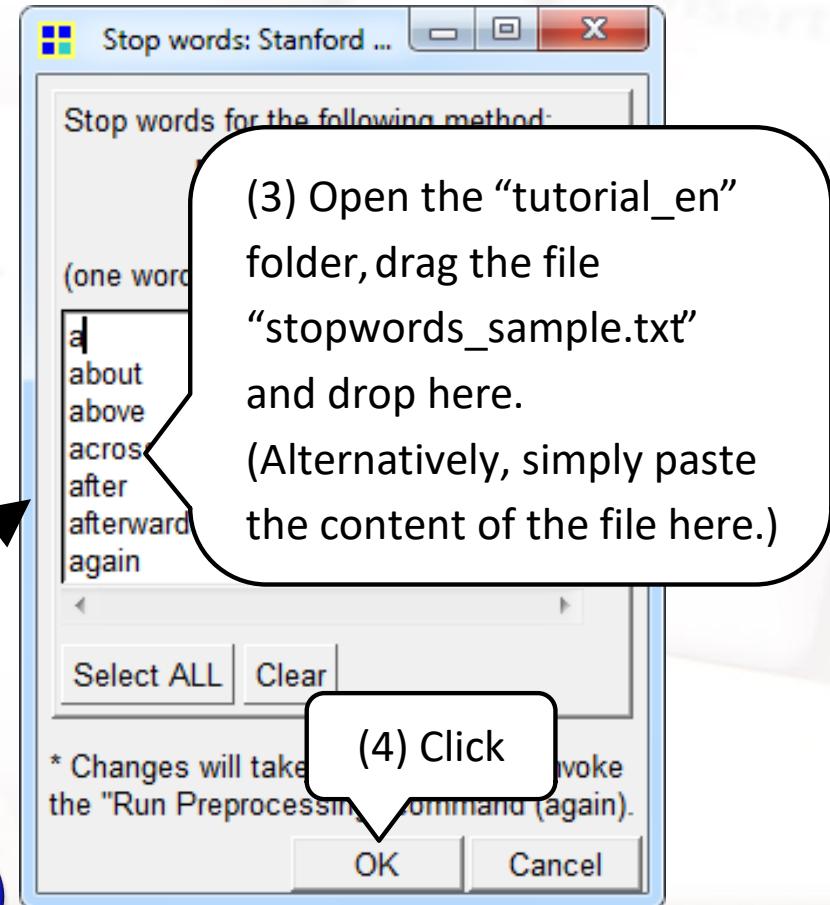
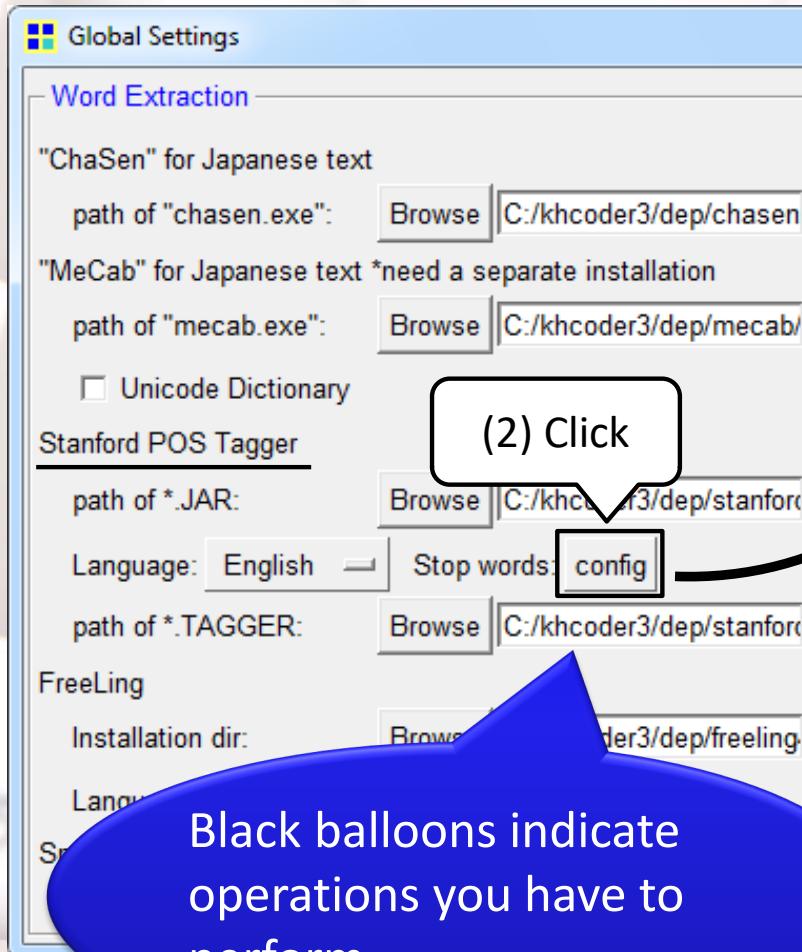
(1) Double click the shortcut on your desktop to start KH Coder



- Interface translation is not completed.
- If you find a typo or if you have a suggestion, post it here:  
<https://sourceforge.net/p/khc/discussion/>

# Configure Stopwords

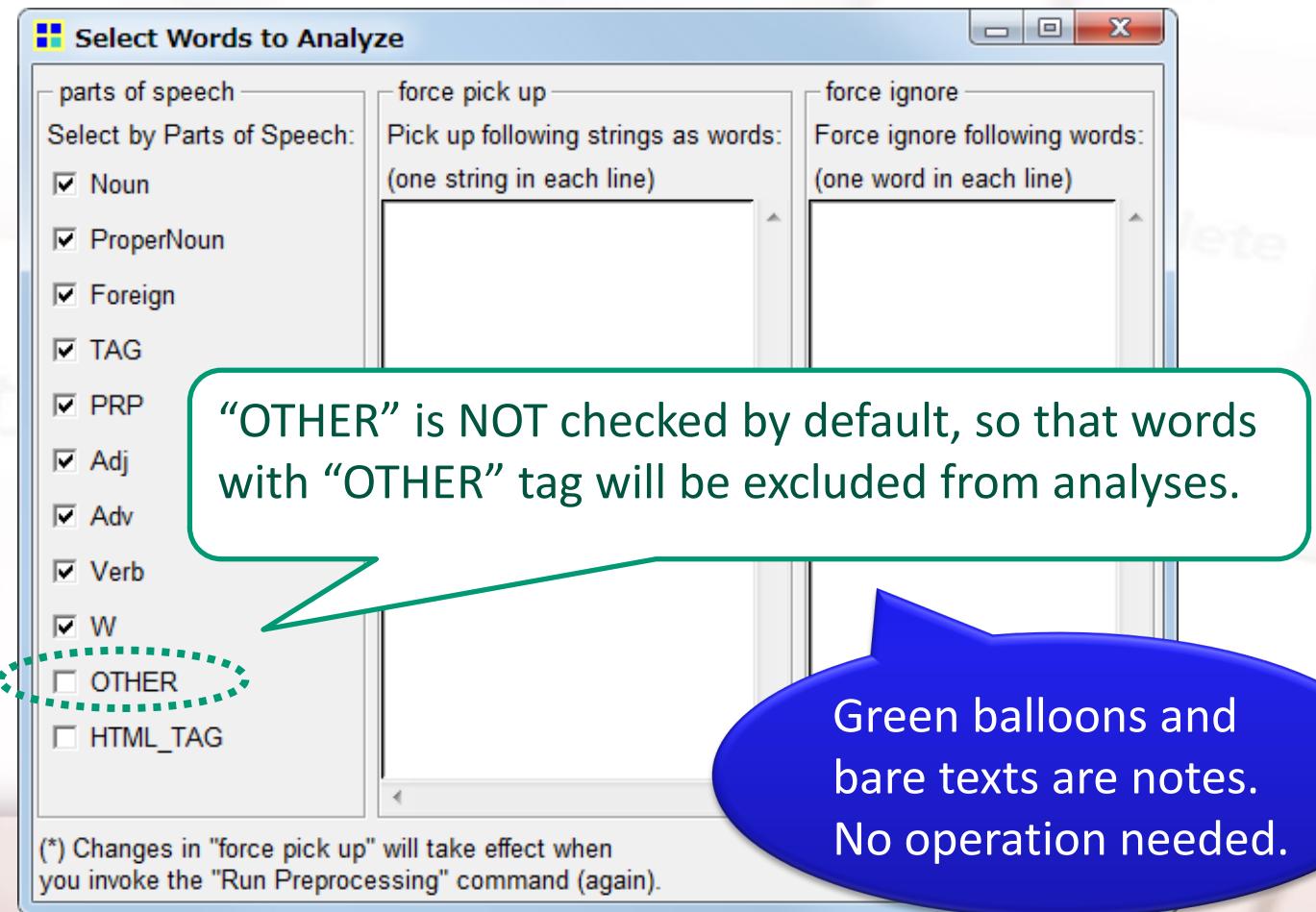
(1) Go to [Project] [Settings] in the menu of KH Coder



Black balloons indicate  
operations you have to  
perform.

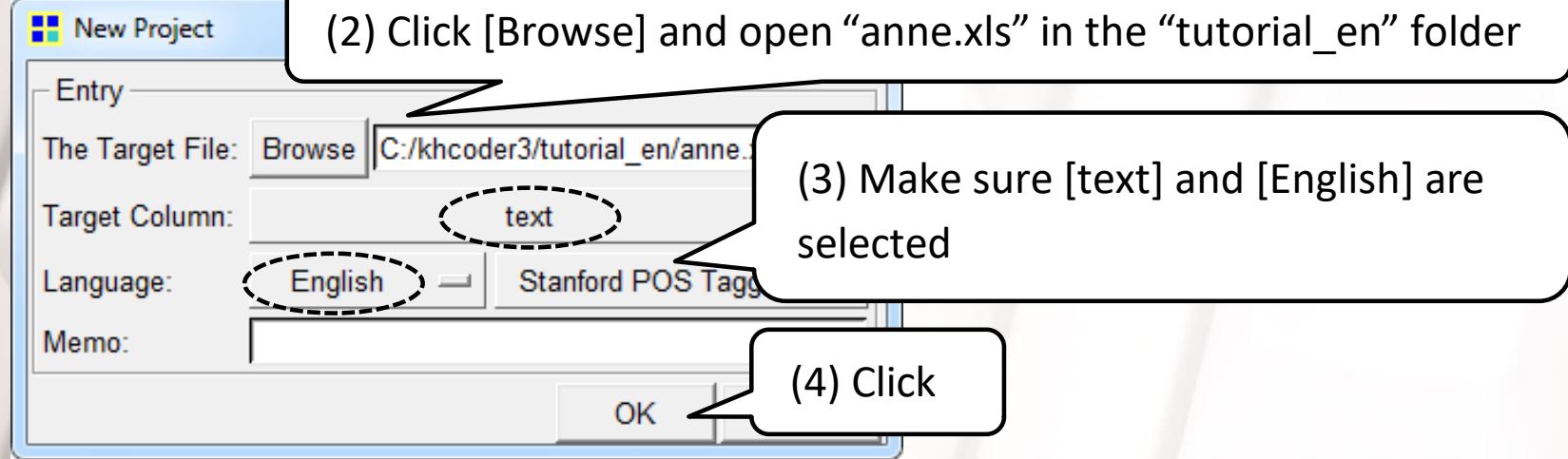
# Notes on Stopwords

- You can specify any words as stopwords in KH Coder to exclude those words from your analysis.
- Stopwords will be given the special POS tag “OTHER”.



# Create a Project & Run Pre-Processing

(1) Go to [Project] [New] in the menu of KH Coder



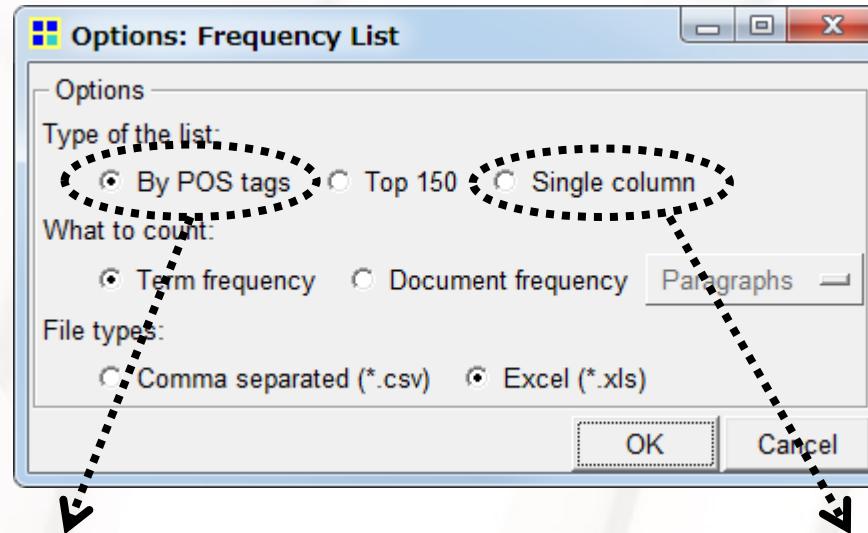
(5) Go to [Pre-Processing] [Run Pre-Processing] in the menu and click [OK]

- Next time you start KH Coder, go to [Project] [Open] in the menu and open the project you have created here.
- When processing data, KH Coder “concentrates” on the task. So it may look frozen or “not responding”. But it is normal when CPU or disk is busy.

# Step 1

# Word Frequency List (1/2)

Go to [Tools] [Words] [Frequency List] in the menu



	A	B	C	D	K	L	M	N
1	Noun		ProperNoun		Adj		Adv	
2	girl	267	ANNE	1138	little	283	just	358
3	thing	260	MARILLA	849	good	225	really	116
4	time	208	Diana	414	old	96	away	97
5	eye	152	Matthew	361	white	91	pretty	79
6	home	136	Lynde	151	..	80	..	77
7	child	134	Barry	115	real	12	far	43
8	school	128	Gilbert	96	sure	70	perfectly	39
9	night	117	Jane	96	..	70	..	39

These numbers are counts of base forms / lemma

	A	B	C
1	Words	POS	TF
2	ANNE	ProperNoun	1138
3	say	Verb	952
4	MARILLA	ProperNoun	849
5	think	Verb	486
6	..	..	414
7	..	..	364
8	..	..	361
9	just	Adv	358

# Word Frequency List (2/2)

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

- The character name that most frequently appears next to the heroine “ANNE” is not her best friend “Diana” but “MARILLA”.
- In the novel, an orphan “girl” or “child” heroine gets adopted, finds a “home”, and goes to “school”. And she once had a inferiority complex about her “hair”.

# The Context Where a Word is Used

(1) Go to [Tools] [Words] [KWIC Concordance] in the menu

The screenshot shows the KWIC Concordance application window. In the search bar, 'Search Entry' is set to 'Word: child'. The result pane displays a paragraph from a document. A callout bubble labeled '(2) Type a word and hit the [Enter] key' points to the search bar. Another callout bubble labeled '(3) Double click a line to view the whole paragraph' points to the paragraph text, specifically highlighting the word 'child' which has been double-clicked.

**KWIC Concordance**

Search Entry: Word: child

POS: Co

Sort 1: None Sort 2: None

Result

Result pane content:

...b...t's about time somebody adopted that child ...t's next door to  
t afternoon. During the forenoon she kept the child busy with various tasks and watched over her  
od little girl and show yourself grateful. Why, child, whatever is the matter? <h5> --cell-- </h5>  
ieve He could really have looked so sad or the children would h...  
ould be tacked on to every remark made to a child who was  
y when you've never had any experience with children. You  
n, I suppose, and there's no guessing how a child like that  
ble skinny and homely, Marilla. Come here, child, and let  
les? And hair as red as carrots! Come here, child, I say.  
mind. You'll have your own troubles with that child. But if you  
se you won't do, although I've brought up ten children and buri...  
e the most effective language for that kind of a child. Her tem...  
ie efficiency of which all of Mrs. Rachel's own children could ha...  
Marilla. She did not believe she could whip a child. No, so  
up before Marilla. She had been a very small child when sh...  
<h5> --cell-- </h5> "There, there, get up, child, " she s...

Document pane content:

<h5> --cell-- </h5>

For reasons best known to herself, Marilla did not tell Anne that she was to stay at Green Gables until the next afternoon. During the forenoon she kept the child busy with various tasks and watched over her with a keen eye while she did them. By noon she had concluded that Anne was smart and obedient, willing to work and quick to learn; her most serious shortcoming seemed to be a tendency to fall into daydreams in the middle of a task and forget all about it until such time as she was sharply recalled to earth by a reprimand or a catastrophe.

\* Search Result: 51 / 134, No. 340

h1 = 0, h2 = 0, h3 = 0, h4 = 0, h5 = 340

# Co-Occurrence Network of Words (1/2)

(1) Go to [Tools] [Words] [Co-Occurrence Network] in the menu

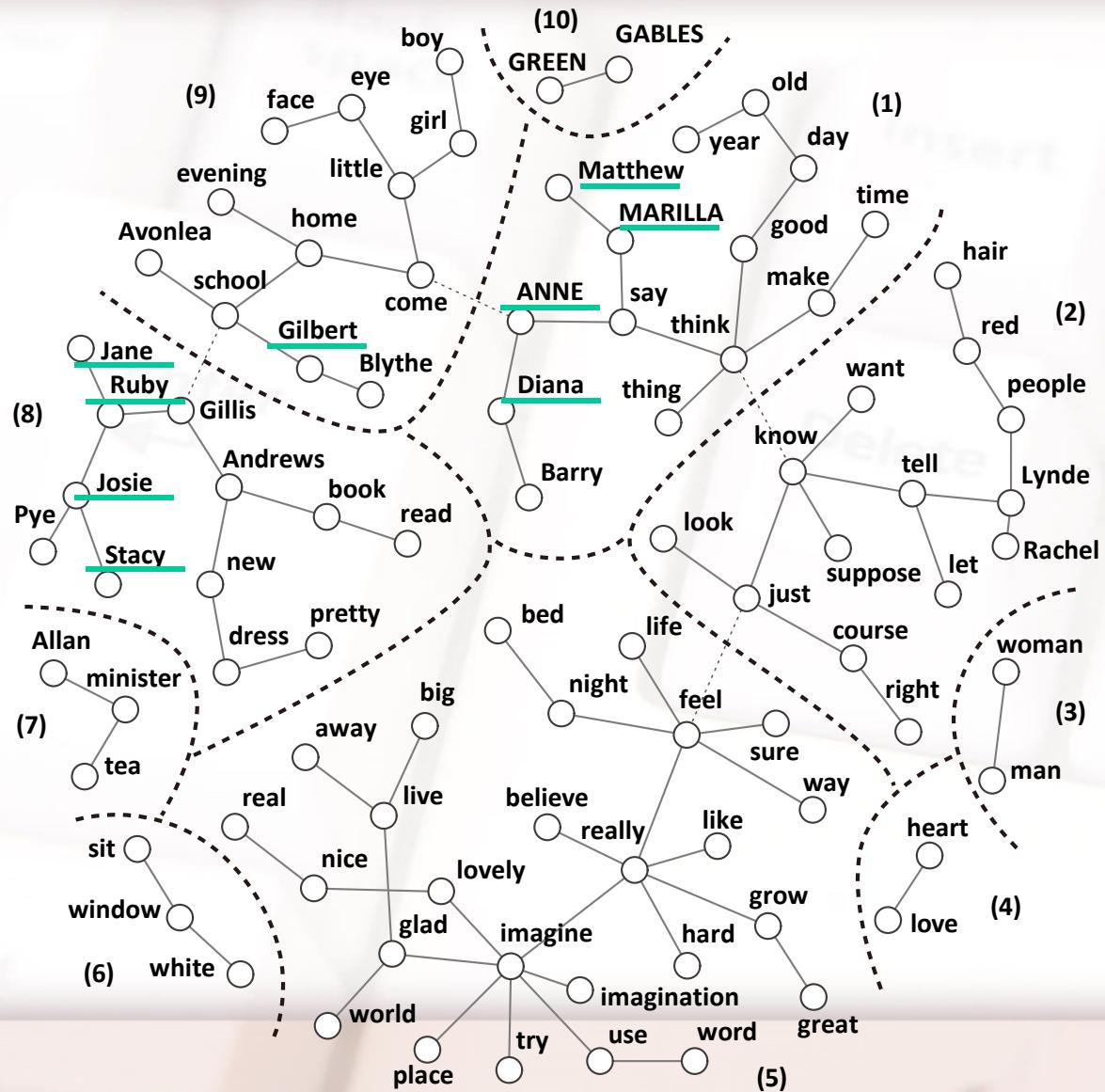
The screenshot shows the 'Options: Co-Occurrence Network of Words' dialog box. On the left, under 'Select the unit & words', the 'Unit' is set to 'H5'. There are three sections for filtering words: 'Filter words by Term Frequency' (Min. TF: 50, Max. TF: 100), 'Filter words by Document Frequency' (Min. DF: 1, Max. DF: 100), and 'Filter words by POS'. A list of checked POS filters includes Noun, ProperNoun, Foreign, TAG, PRP, Adj, Adv, and Verb. Below this is a 'Number of selected words' field with 'Check' and '123'. On the right, the 'Co-Occurrence Network Options' panel has 'Type of edges' set to 'words - words' and 'Variable / Headings' set to 'Heading 5'. It includes several configuration checkboxes: 'Top 240' (selected), 'Jaccard coef. >= 0.2', 'Thicker lines for stronger edges', 'Show coef.', 'Larger nodes for higher frequency words', 'Variable font size \*For printing with EMF/EPS/PDF' (unchecked), 'Smaller Nodes' (selected), 'Highlight the minimum spanning tree', 'Draw the minimum spanning tree only' (selected), 'Avoid overlapping of labels', 'Translucent Colors (not suitable for EMF/EPS)' (unchecked), and 'Grayscale (centralities & communities)' (unchecked). At the bottom are 'Plot size: 640', 'OK', and 'Cancel' buttons. To the right of the dialog is a network visualization window titled 'Co-Occurrence Network of Words' showing a graph of words like boy, girl, eye, home, etc., connected by lines of varying thickness. A callout bubble points to the 'Communities: modularity' checkbox in the network window.

(2) Configure as shown in this screen and click [OK]

(3) Select [Communities: modularity] here

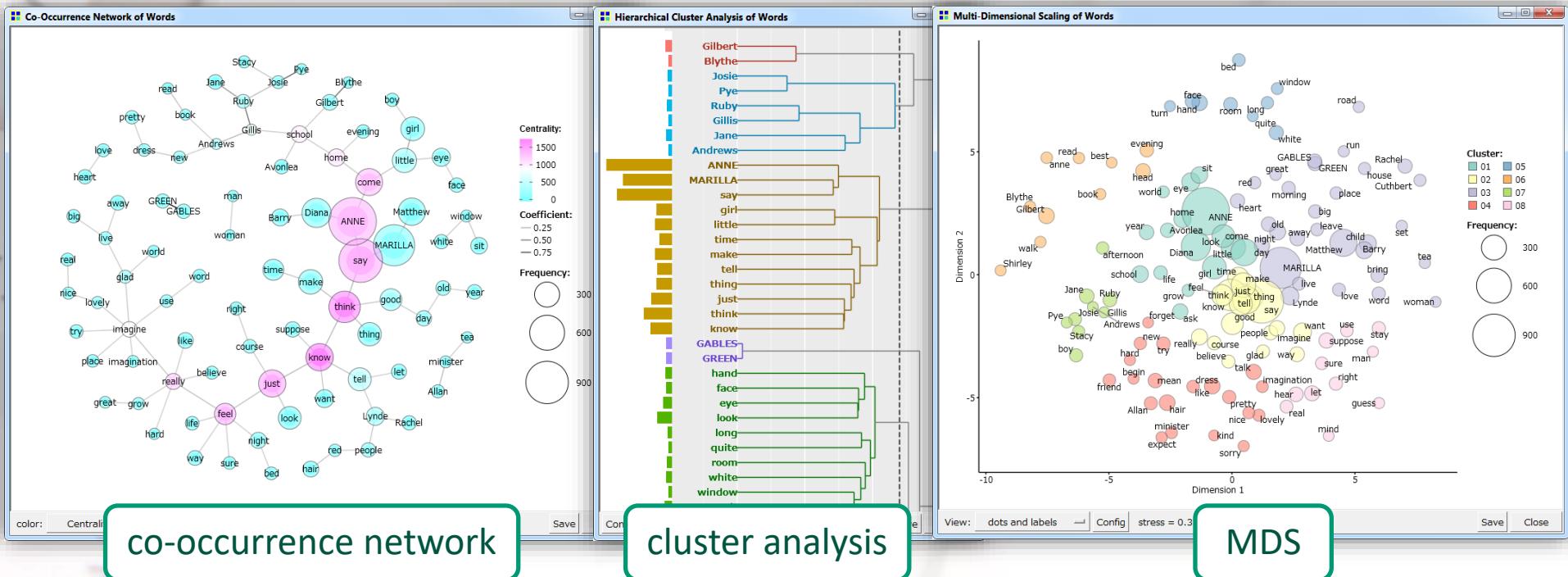
# Co-Occurrence Network of Words (2/2)

- “Diana”, “Marilla”, and “Matthew” are connected close to “Anne”
- “Gilbert” is in rather remote part and connected to “Anne” via “school”
- “Jane”, “Ruby”, “Josie”, and “Stacy” are also connected via “school”
- The figure is retouched with Illustrator



# Methods for Exploring Co-Occurrences of Words

- To explore co-occurrences of words, you can also use:
  - ✓ hierarchical cluster analysis
  - ✓ Multi-dimensional scaling



co-occurrence network

cluster analysis

MDS

- By interpreting these results, you may find major themes of the text from groups of words which tend to appear together.
- KH Coder uses R as back end to execute these multivariate methods.

# Correspondence Analysis of Words (1/2)

(1) Go to [Tools] [Words] [Correspondence Analysis] in the menu

**Options: Correspondence Analysis of Words**

Select Words

Filter words by Term Frequency  
Min. TF: 50 Max. TF:

Filter words by Document Frequency  
Min. DF: 1 Max. DF:   
Document Unit: H5

Filter words by POS

- Noun
- ProperNoun
- Foreign
- TAG
- PRP
- Adj
- Adv
- Verb
- W

All Default Clear

Number of selected words  
Check 122

Correspondence Analysis Options

Input Data Matrix:

Words x Documents  
Tabulating Unit: H5  
 Biplot

Words x Variable(s)  
part  
chapter

Filter words by chi-square value: Top 40

Show labels only for distinctive words: Top 60

Bubble plot: Size of bubbles 100 %  
 Size of variables also reflect word counts  
 Translucent Colors (not suitable for EMF/EPS)

Dimensions of the Plot: X 1 Y 2

Scaling: none  Show the origin  
Dot size: 640

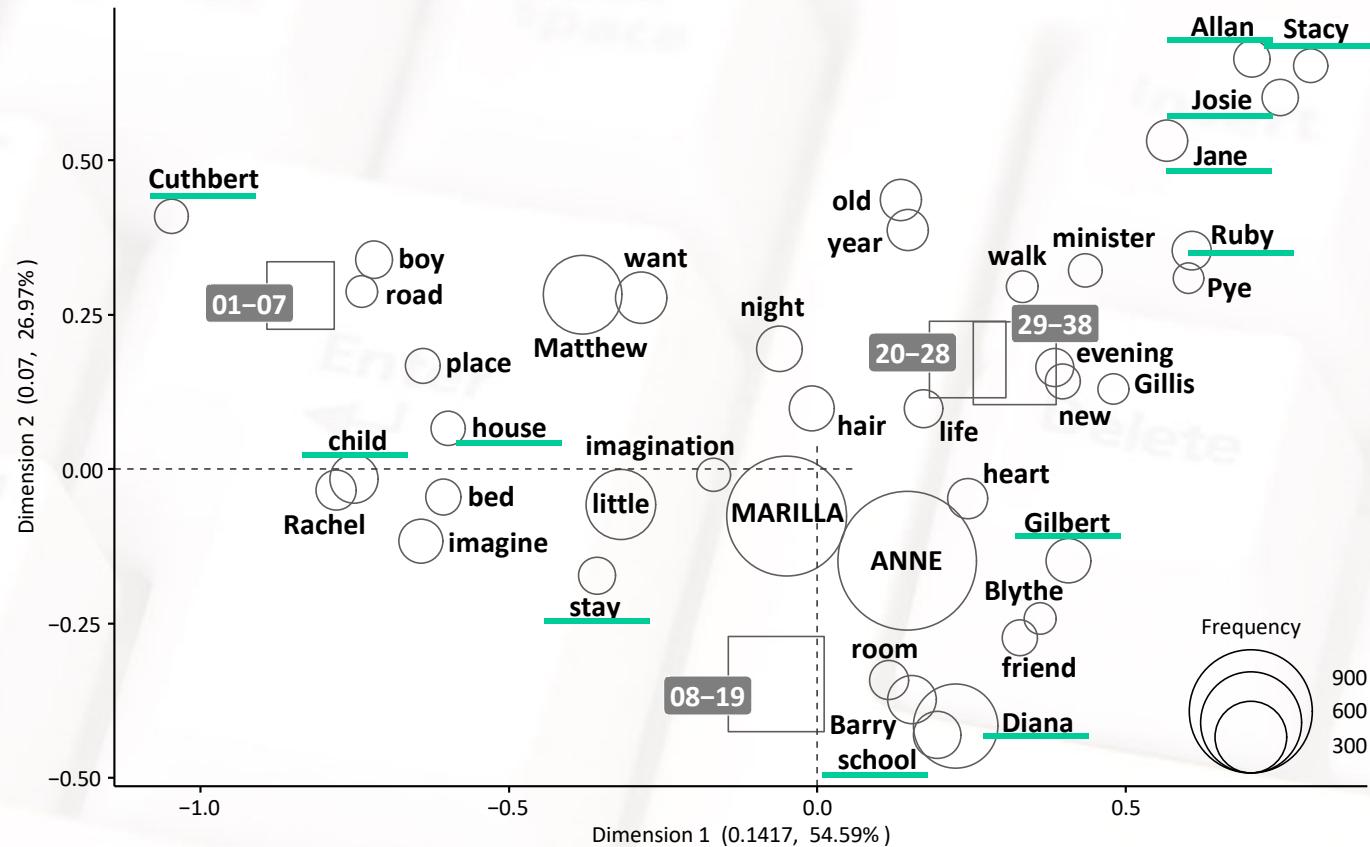
View: color  Save Close

**(2) Configure as shown in this screen and click [OK]**

**(3) Select [grayscale] here**

# Correspondence Analysis of Words (2/2)

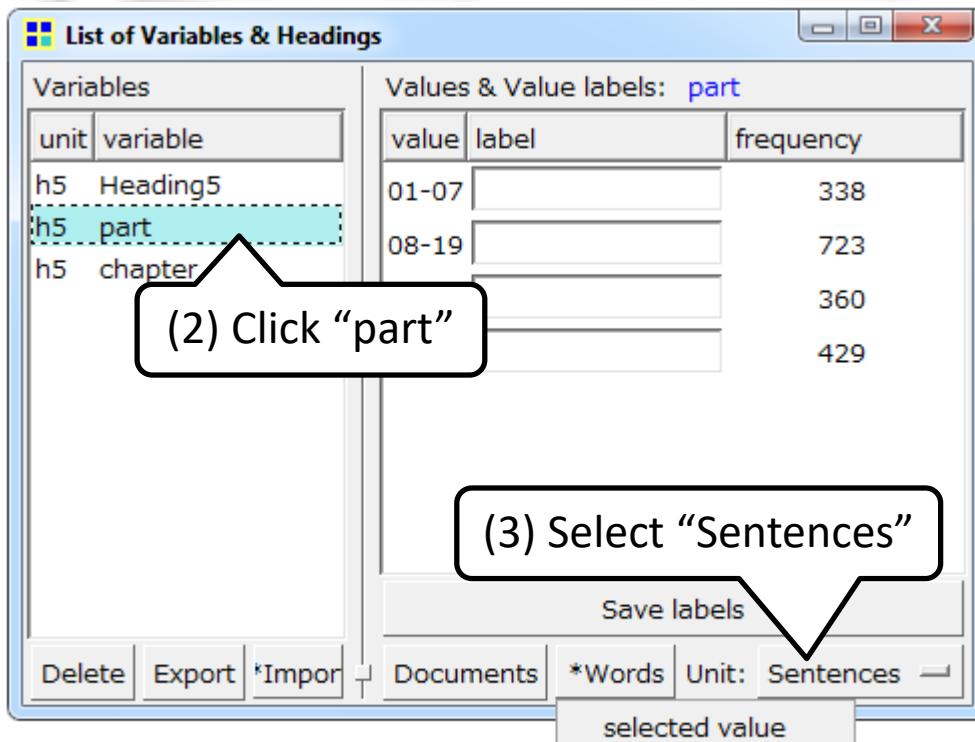
- In the beginning [01-07], the “child” Anne was allowed to “stay” in “Cuthbert’s house”.
- Then in [08-19], she met a neighbor girl “Diana” and started going to “school”. At the school, she met “Gilbert”.
- In the latter half of the novel, Anne and Diana went separate ways, and Anne's schoolmates, such as “Josie”, “Jane”, and “Ruby”, become characteristic. Anne also learned a lot from adult women such as Mrs. “Allan” and Miss “Stacy”.



We can understand the story flow throughout the novel by checking characteristic words of each part.

# Characteristic Words of each Part

(1) Go to [Tools] [Variables & Headings] in the menu



	01-07	08-19	20-28
say	.087	ANNE	.151
Matthew	.075	MARILLA	.114
little	.045	Diana	.085
come	.045	just	.053
know	.042	little	.043
child	.038	tell	.039
thing	.036	school	.028
look	.033	Barry	.027
girl	.033	Lynde	.025
Spencer	.032	child	.022
			.034

Top 10 characteristic words of each part are tabulated. It can be used as an alternative for correspondence analysis.

# Closing Remarks for Step 1

▣ Statistical analyses of automatically extracted words are suitable for gaining a whole picture of the data

- ✓ Main theme (word frequency list or co-occurrence network)
- ✓ Relations between characters or words (co-occurrence network)
- ✓ Story flow (correspondence analysis)

▣ About the centrality of Marilla

- ✓ Most frequently appears next to the heroine Anne
- ✓ Her relationship with Anne appears to be almost as strong as Diana's
- ✓ Be present throughout all four parts of the story

We obtained overviews of entire data in this step. Next, we are going to put more focus on Marilla using coding rules.

# Step 2

# Use Coding Rules to Count Concepts

- In some cases, we have to count concepts, not words.
- To count concepts, you can compose “coding rules” like this:

Indicates the name of this code: “Character\_name\_Gilbert”

\*Character\_name\_Gilbert

Gilbert or Gil

Not only the documents containing “Gilbert” but  
also those containing “Gil” are assigned this code.

- If a document is acceptable under multiple coding rules, multiple codes will be assigned to the document.

# Retrieve Documents Assigned a Specific Code

(1) Go to [Tools] [Documents] [Search Documents] in the menu

(2) Click [Browse] and open "code\_1.txt"

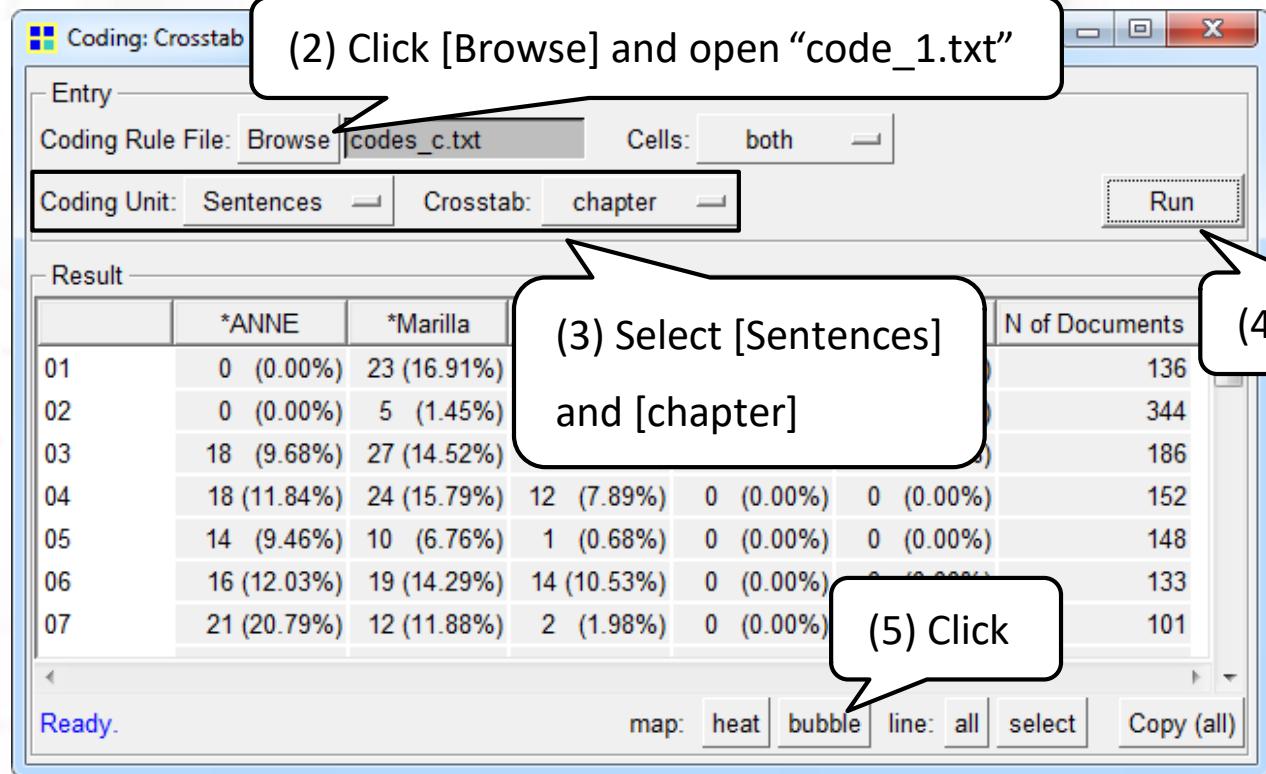
(3) Double click any of the codes

(4) Double click a line to view the whole paragraph

The screenshot shows the 'Search Doc' dialog box. In the 'Search Entry' list, the code '\*ANNE' is selected. The 'Coding Rule File' dropdown shows 'code\_1.txt'. Below the list, the search criteria are set to '#direct and AND no sort Unit: Paragraphs Run'. The 'Result:' pane displays a portion of a text document with several lines highlighted in blue, indicating they contain the searched code. A callout box labeled '(4)' points to one of these highlighted lines. A separate 'Document' window is also visible, showing the full context of the highlighted line.

# Characters in Each Chapter (1/2)

(1) Go to [Tools] [Coding] [Crosstab] in the menu



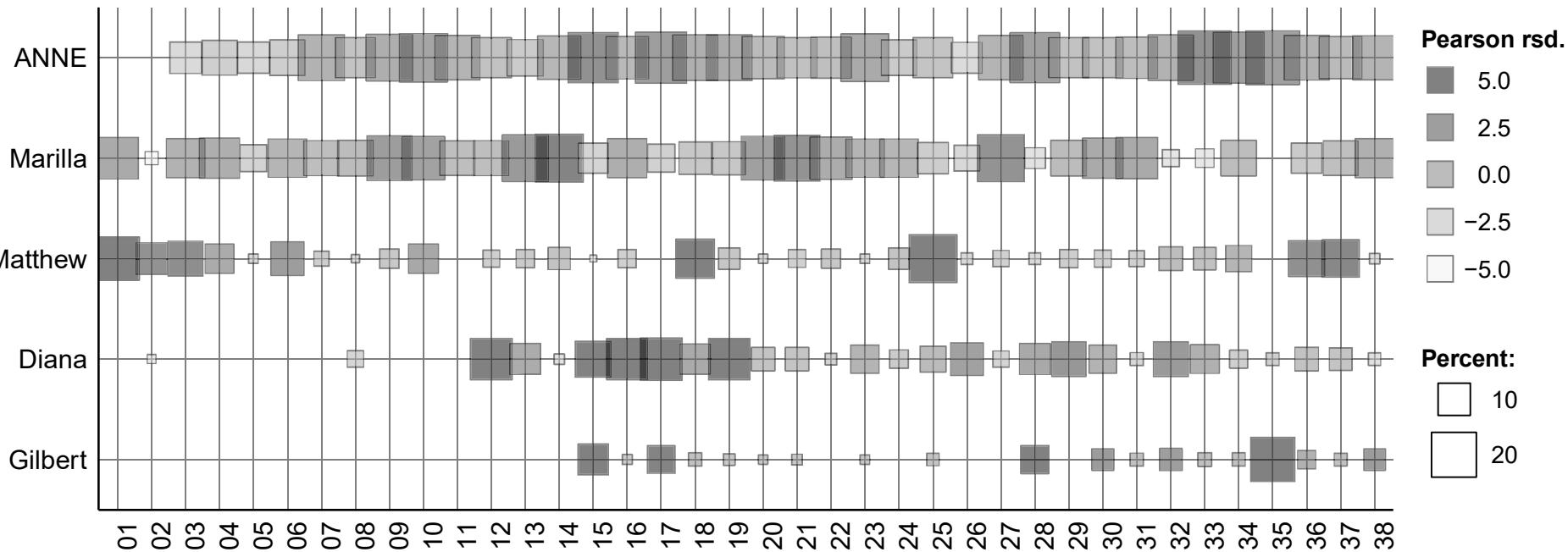
(2) Click [Browse] and open “code\_1.txt”

(3) Select [Sentences] and [chapter]

(4) Click

(5) Click

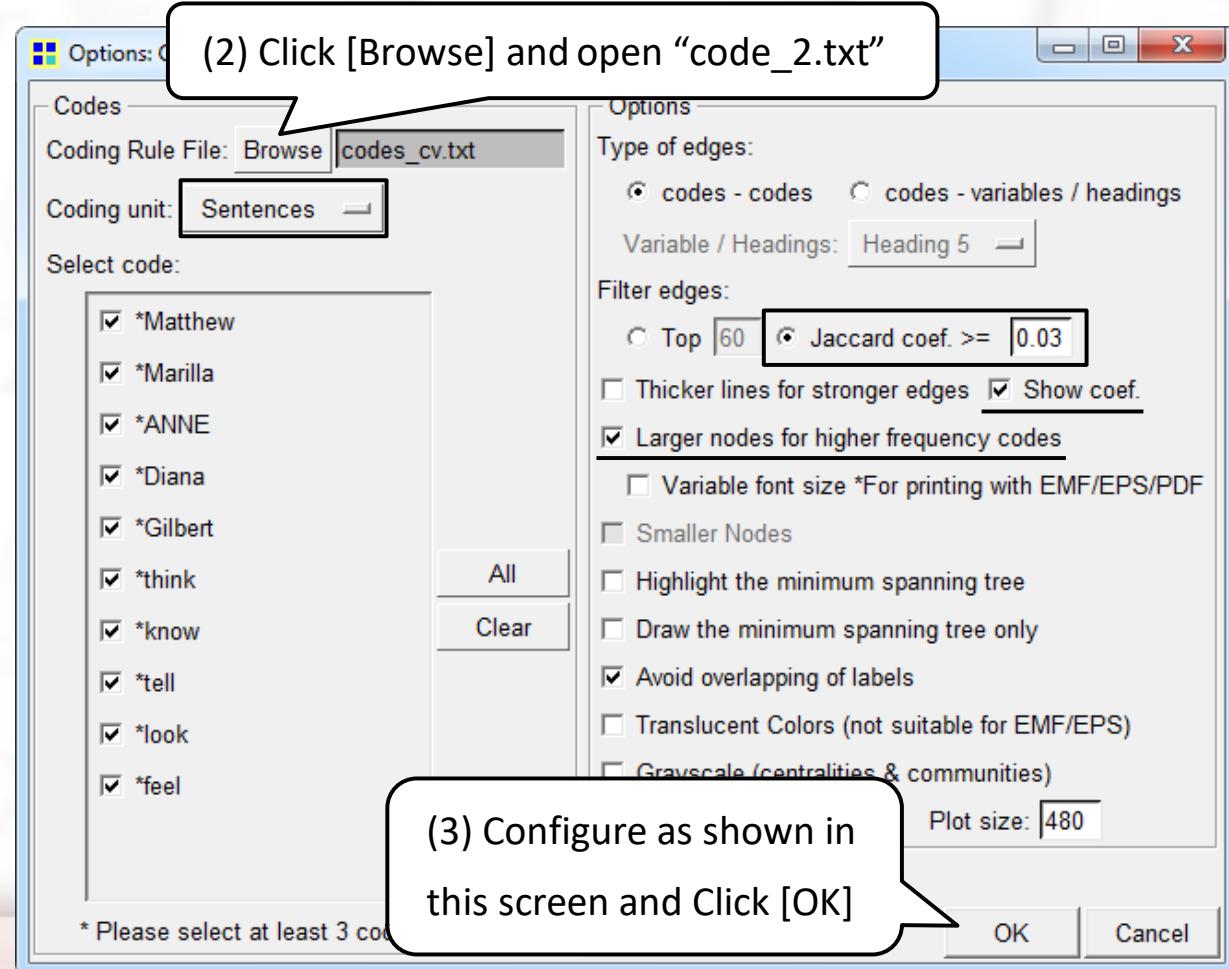
# Characters in Each Chapter (2/2)



- Marilla and Anne are present almost everywhere
- Although Marilla and Anne were apart in chapter 35, there was an emotional reunion in the following chapter 36. Anne won a scholarship and rejoiced saying “Oh, won’t Matthew and Marilla be pleased!”

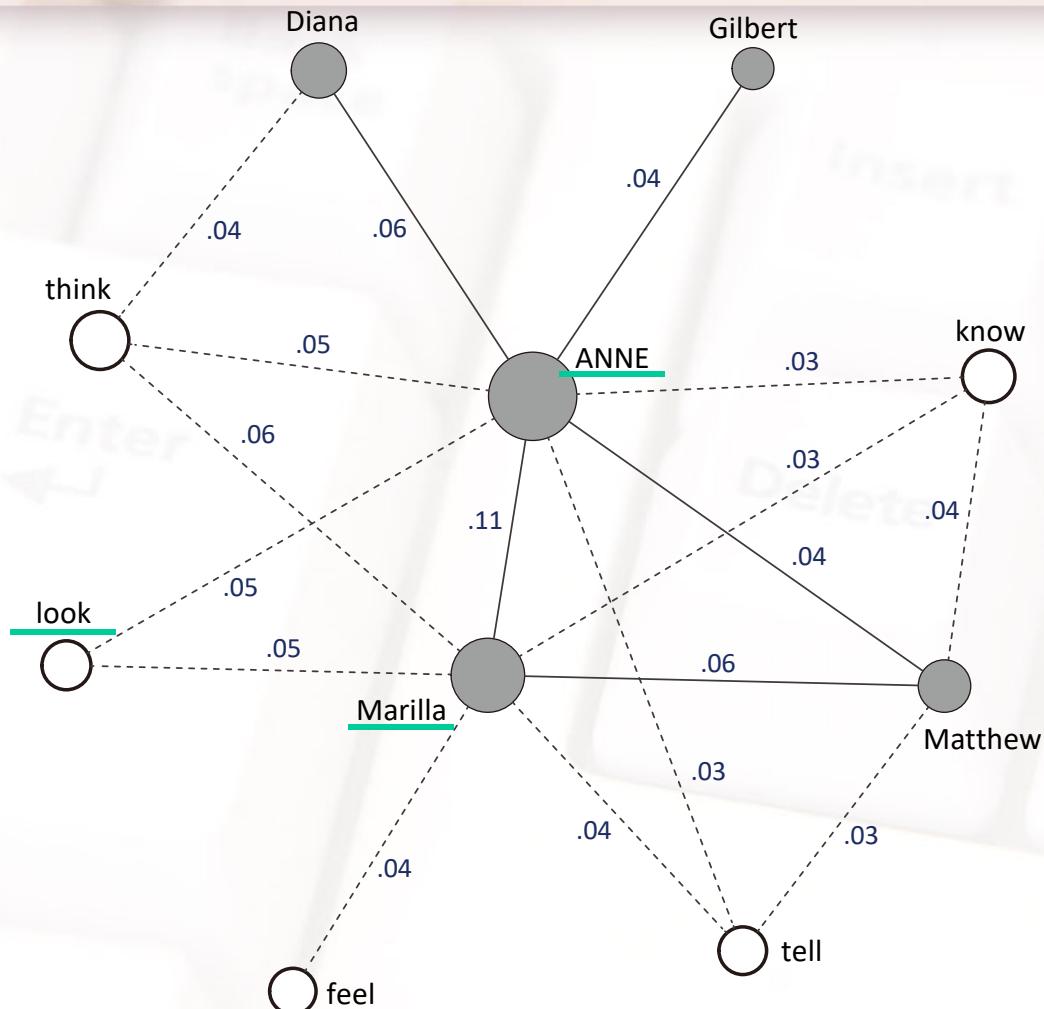
# Characters and Verbs (1/2)

(1) Go to [Tools] [Coding] [Co-occurrences Network] in the menu



# Characters and Verbs (2/2)

- Anne often expresses what she “feels” to Marilla:
  - ✓ “I do feel dreadfully sad, Marilla” (c21)
  
- Marilla and Anne often “look” at each other:
  - ✓ Marilla looked at Anne and softened at sight of the child’s pale face... (c6)
  - ✓ Anne looked at her with eyes limpid with sympathy (c20)
  - ✓ Marilla looked at her with a tenderness that would never have been suffered to reveal itself in any clearer light... (c30)



Marilla and Anne exchange their feelings by words, and also with their eyes, meaning that a close and intimate relationship is depicted between the two.

# Change of Words Co-occurring with Marilla (1/3)

(1) Go to [Tools] [Words] [Word Association] in the menu

(3) Click [\*Marilla]

(2) Click [Browse] and open "code\_3.txt"

(4) Hold down [Ctrl]  
key on the keyboard  
and click [\*01-07]

(5) Click

3 Cuthbert	ProperNoun	64 (0.009)	7 (0.058)	0.0395
4 table	Noun	43 (0.006)	6 (0.050)	0.0382
5 dish	Noun	20 (0.003)	5 (0.042)	0.0370
6 child	Noun	132 (0.019)	8 (0.067)	0.0328
7 bed	Noun	71 (0.010)	6 (0.050)	
8 say	Verb	902 (0.133)	32 (0.267)	
9 uncomfortable	Adj	9 (0.001)	4 (0.033)	
10 sorrel	Noun	11 (0.002)	4 (0.033)	

Copy KWIC Sort: Jaccard Filter

\* To search the words co-occurring with Marilla in the following part "08-19", repeat procedure (3) and then click [\*08-19] instead of [\*01-07] in procedure (4).

# Change of Words Co-occurring with Marilla (2/3)

"Marilla really did not know how to talk to the child, and her uncomfortable ignorance made her crisp and..." (c4)

The "feel" and "look"

	01-07	08-19		20-28		29-38	
Matthew	.053	<u>say</u>	.072	<u>say</u>	.042	Matthew	.041
mare	.040	<u>ANNE</u>	.059	think	.034	<u>look</u>	.040
Cuthbert	.040	just	.039	<u>ANNE</u>	.032	sit	.039
table	.038	think	.036	cake	.030	ANNE	.038
dish	.037	brooch	.031	make	.028	say	.038
<u>child</u>	<u>.033</u>	tell	.030	minister	.028	face	.031
bed	.032	evening	.025	Allan	.026	girl	.026
say	.032	home	.024	<u>feel</u>	<u>.025</u>	think	.024
<u>uncomfortable</u>	<u>.032</u>	set	.024	know	.024	want	.022
sorrel	.032	let	.023	time	.023	lean	.022

The "child" is upgraded to "Anne" and implying that it is impossible to bring up a child without "saying" anything.

# Change of Words Co-occurring with Marilla (3/3)



## Change of Marilla

1. Uncomfortable ignorance [01-07]



2. Calling Anne and Saying many things [08-28]



3. Exchanging feelings by words and eyes  
with Anne [20-38]

The change is depicted throughout the story.

# Conclusions



Results of step 2 showed that:

- ✓ Marilla is literally present almost everywhere
- ✓ A close and intimate relationship is depicted between Marilla and Anne
- ✓ Change of Marilla and growing relationship between Marilla and Anne is depicted throughout the story

Our analysis supports the assertion that Marilla plays central roll in the story.

Identifying keywords like “child”, “uncomfortable”, “look”, and “feel” through quantitative analysis is considered to be useful for extracting depiction which specifically describes Marilla’s roll and change in the story.



## Web site of KH Coder

<http://khc.sourceforge.net/en>



## For more details on this tutorial

Part 1: <http://www.ritsumei.ac.jp/file.jsp?id=325881>

Part 2: <http://www.ritsumei.ac.jp/file.jsp?id=346128>



## Questions or Comments?

<https://sourceforge.net/p/khc/discussion/>