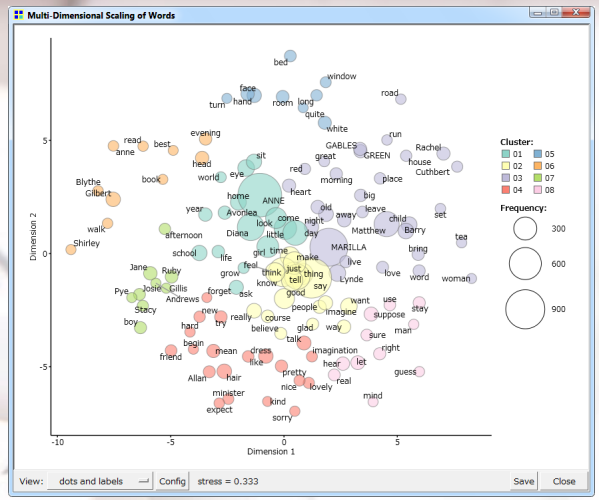
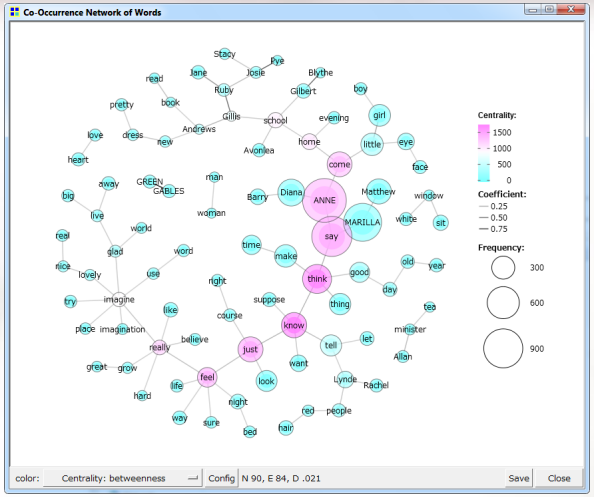
KH Coder Tutorial using *Anne of Green Gables*:

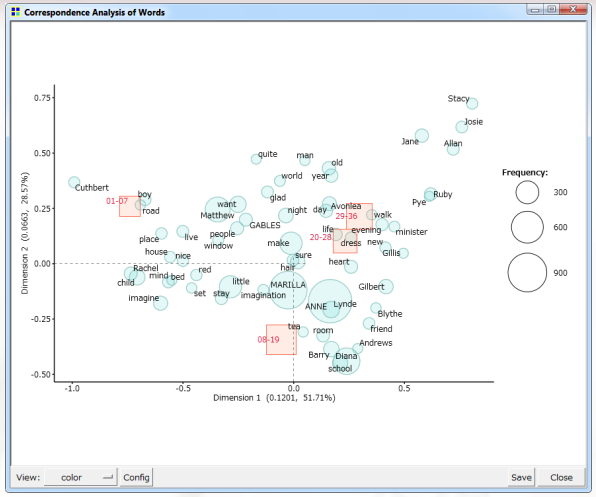
A Two-Step Approach to Quantitative Content Analysis

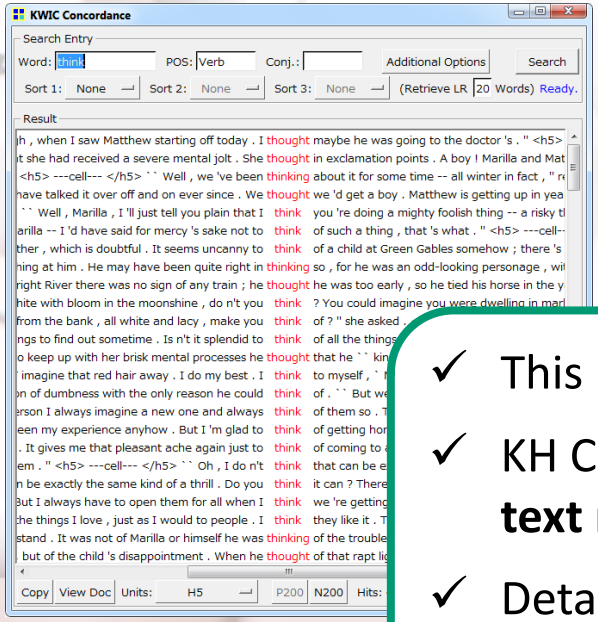
Koichi Higuchi

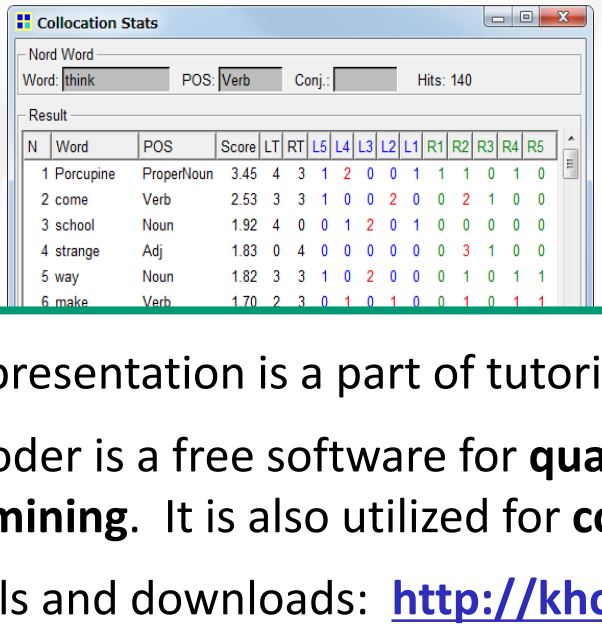
Preface

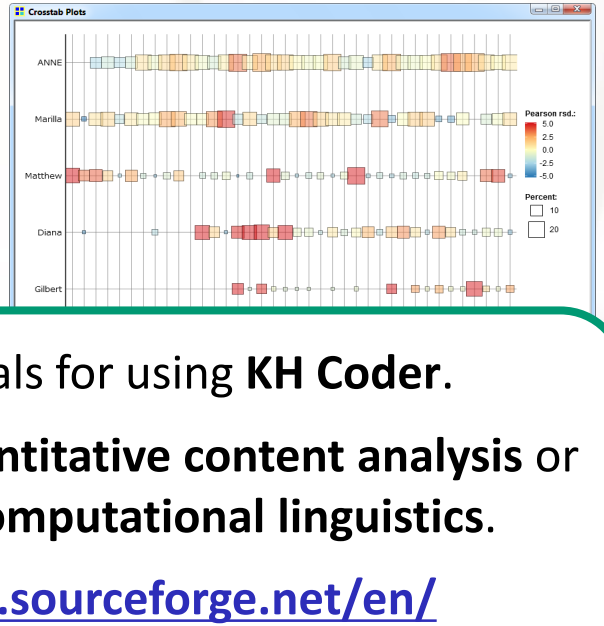












- ✓ This presentation is a part of tutorials for using **KH Coder**.
- ✓ KH Coder is a free software for **quantitative content analysis** or **text mining**. It is also utilized for **computational linguistics**.
- ✓ Details and downloads: http://khc.sourceforge.net/en/

Table of Contents

Introduction

Data

- 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.

(1) Enter column names in the first row

	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 rc	01	01-07
4	There are plenty o 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 t Gables after tea and find ou	01	01-07
9	Acc		-07
10	"It's just STAYING, that's what," she said as she steppe	01	01-07

(2) Enter actual data in the second and subsequent rows

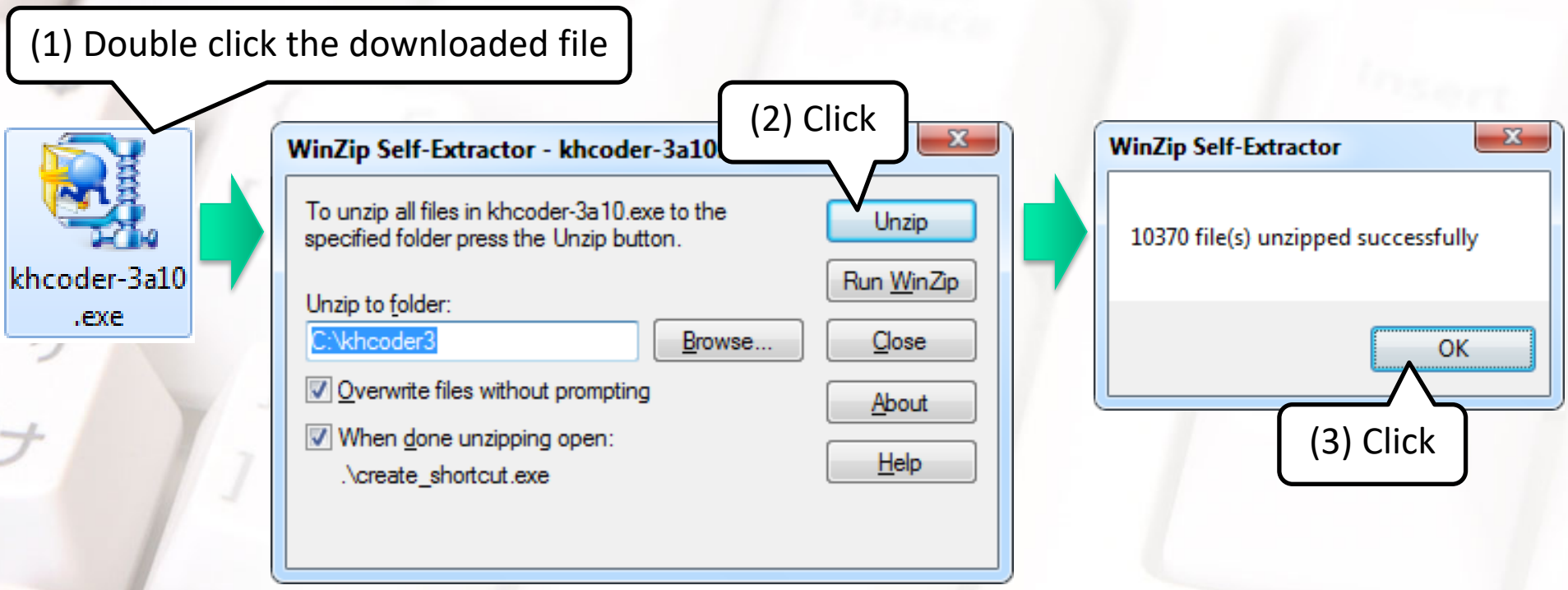
(*) Enter data in the first sheet if you use Excel or Calc

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

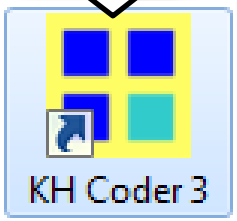
Install KH Coder



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

Interface Language

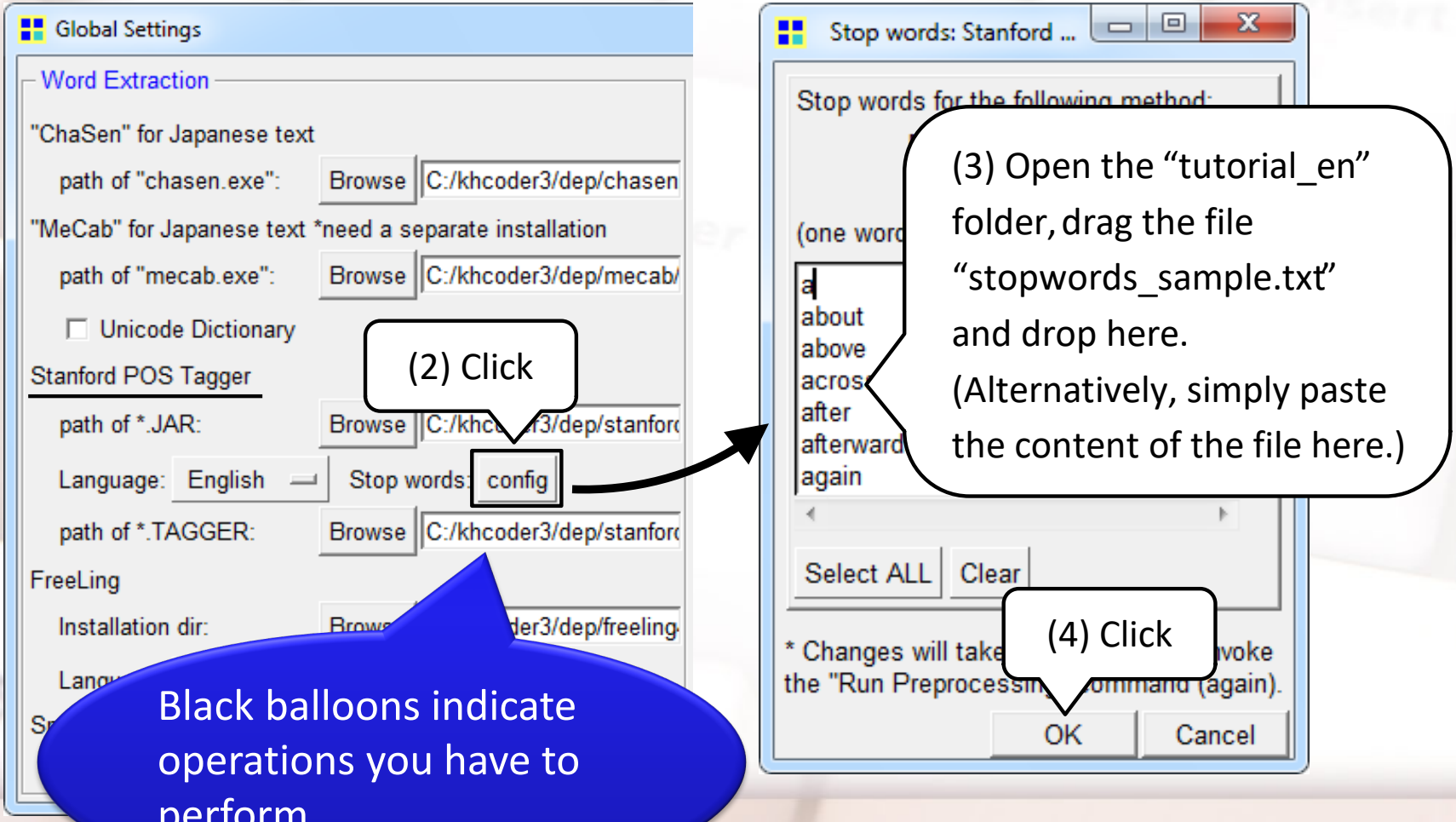
(1) Double click the shortcut on your desktop to start KH Coder



- 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/

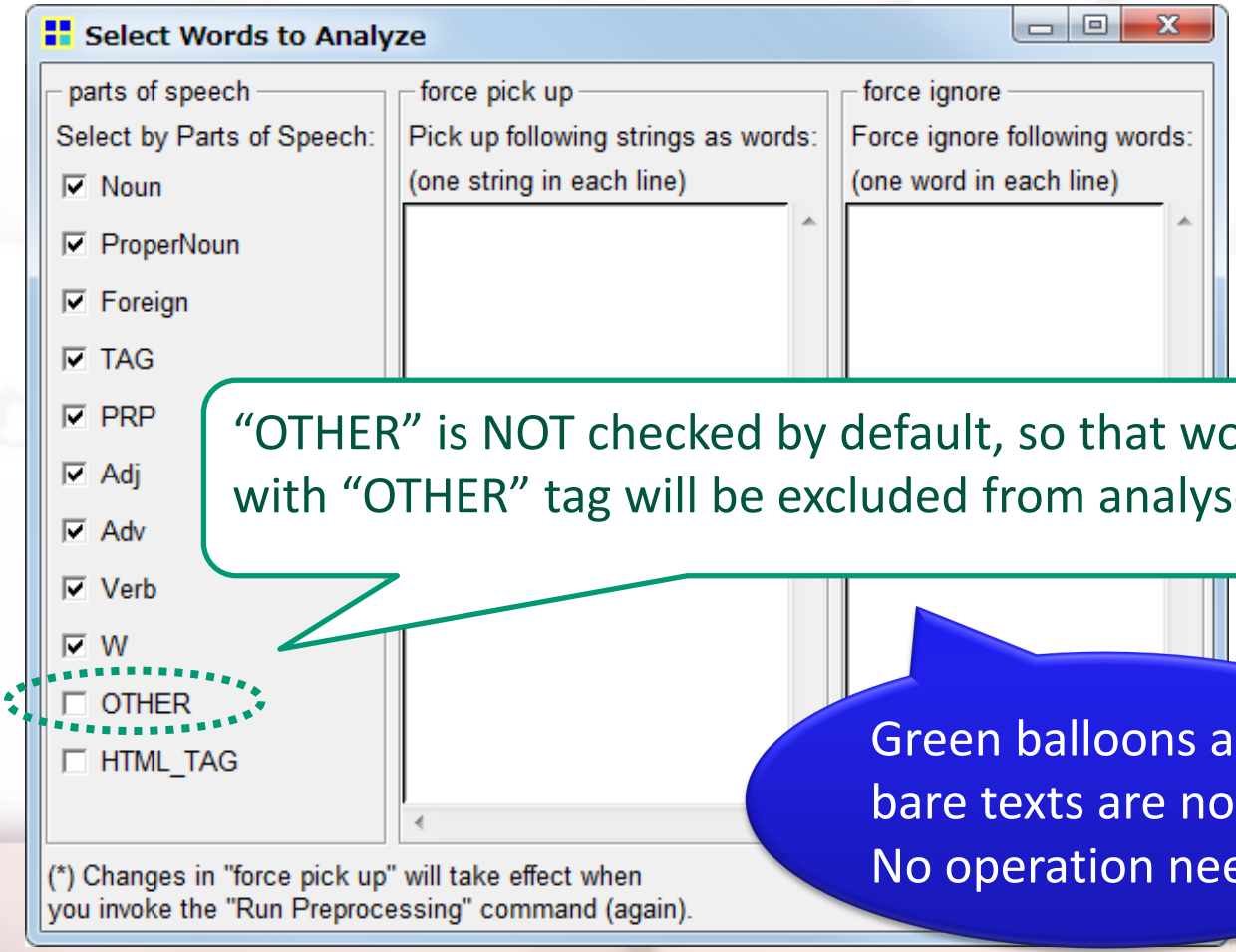
Configure Stopwords

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



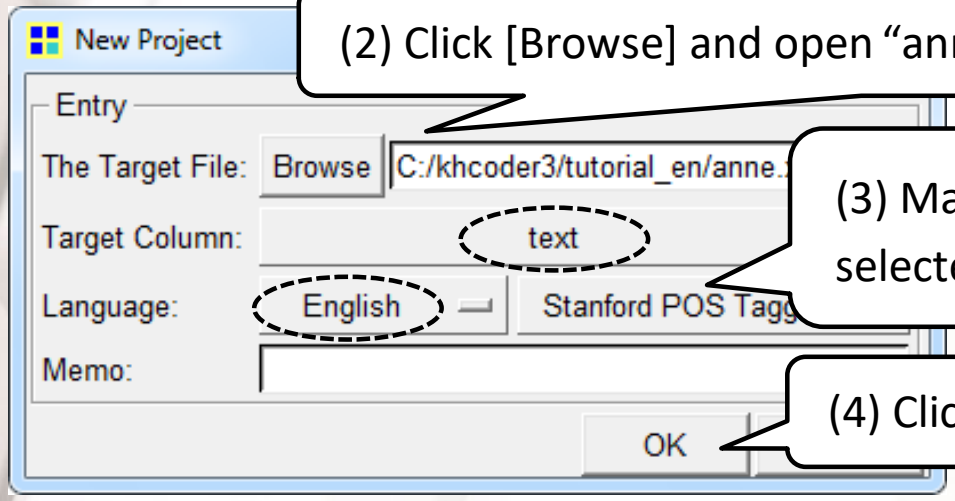
Notes on Stopwords

- 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



(2) Click [Browse] and open “anne.xls” in the “tutorial_en” folder

(3) Make sure [text] and [English] are selected

(4) Click

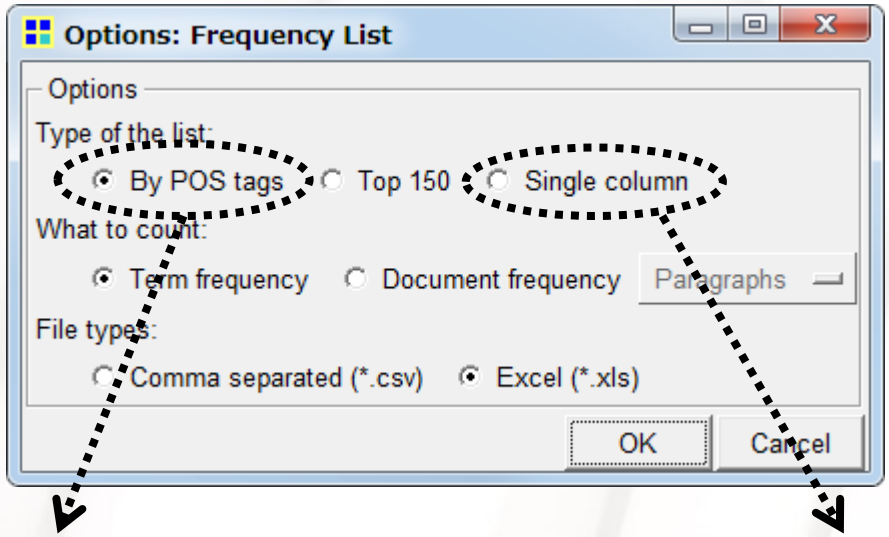
(5) Go to [Pre-Processing] [Run Pre-Processing] in the menu and click [OK]

- 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					
7	child	134	Barry					
8	school	128	Gilbert					
9	night	117	Jane	96	sure	70	perfectly	39

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

These numbers are counts of base forms / lemma

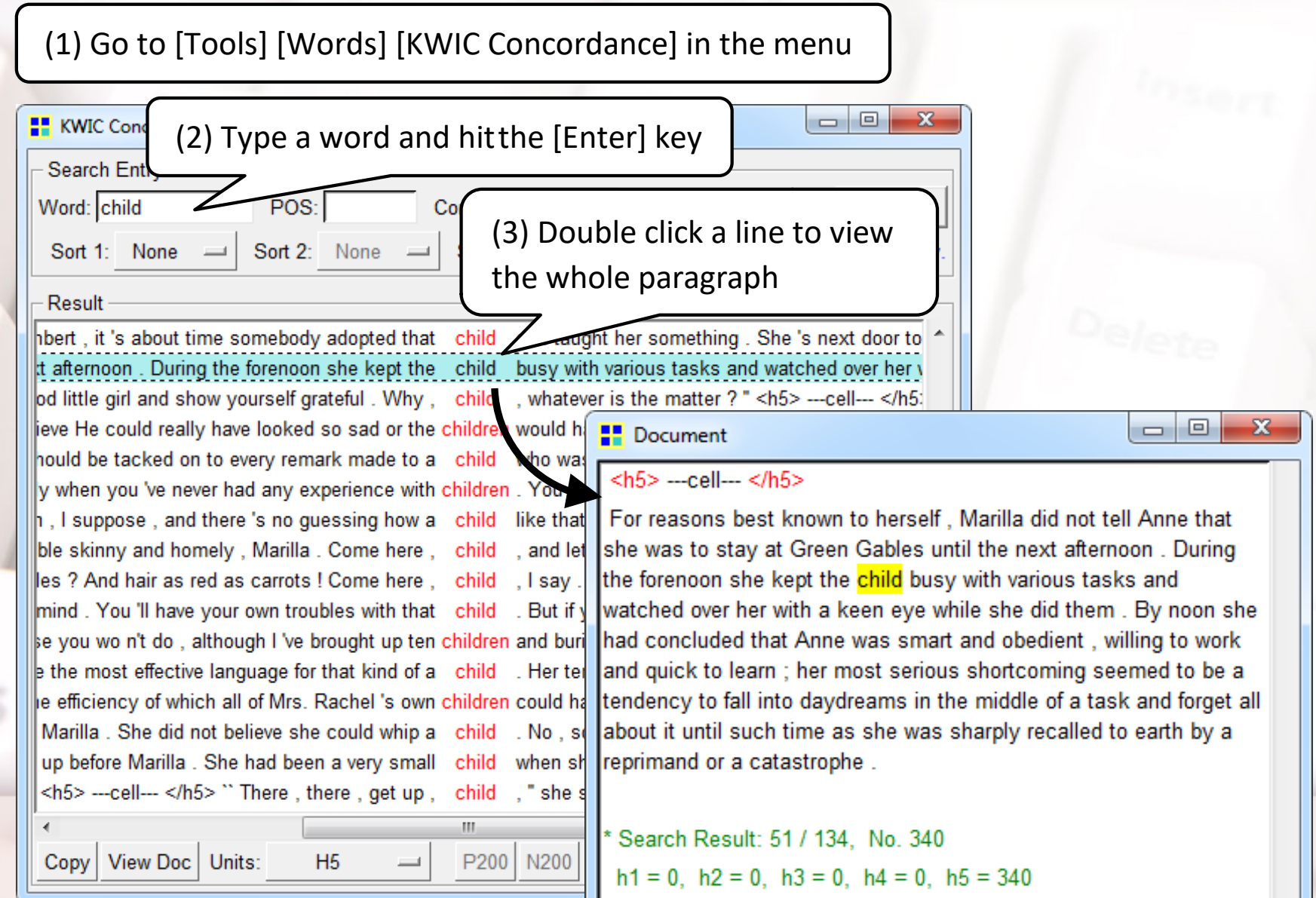
Word Frequency List (2/2)

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

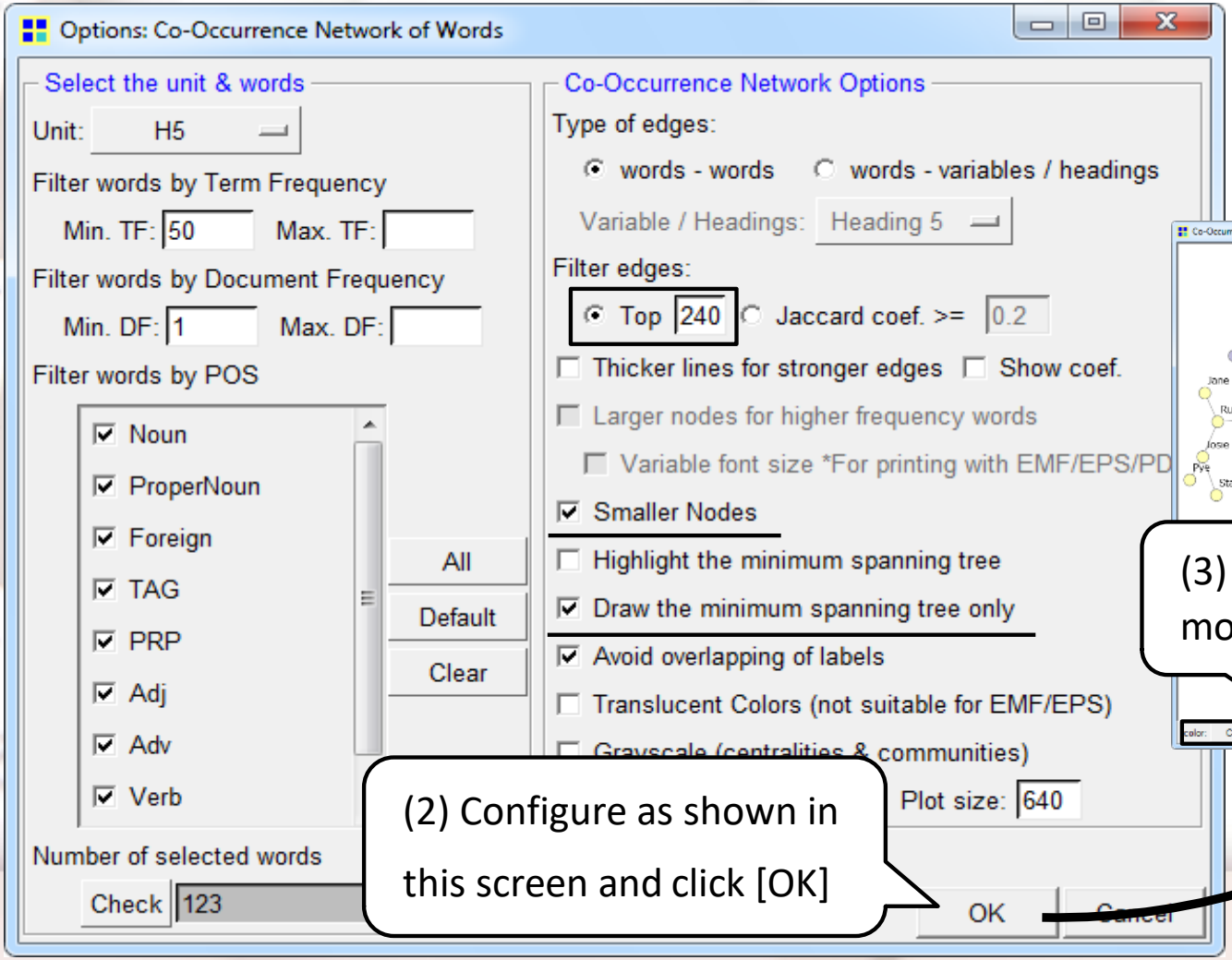
(1) Go to [Tools] [Words] [KWIC Concordance] in the menu



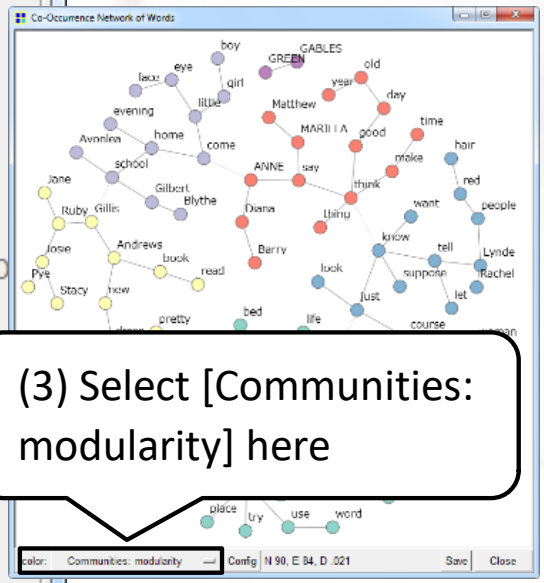
Step 1

Co-Occurrence Network of Words (1/2)

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



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

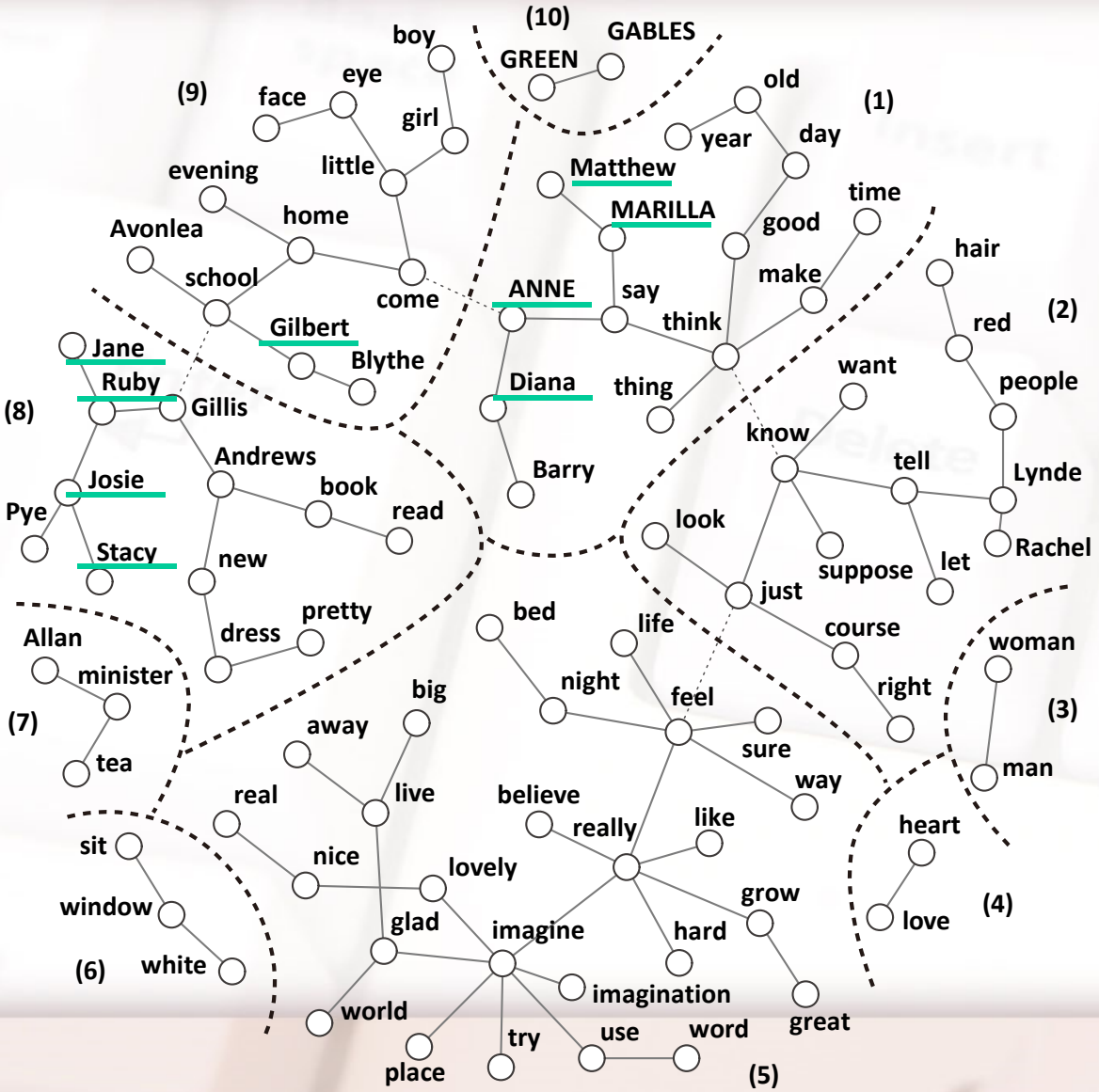


(3) Select [Communities: modularity] here

Step 1

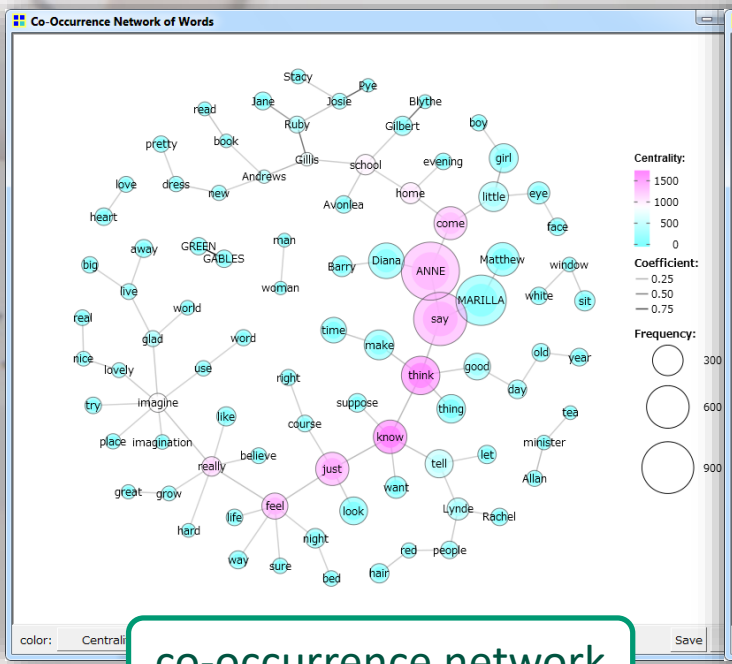
Co-Occurrence Network of Words (2/2)

- “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

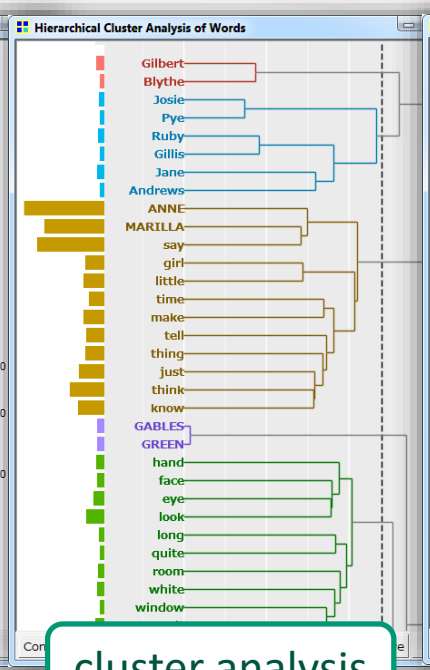


Methods for Exploring Co-Occurrences of Words

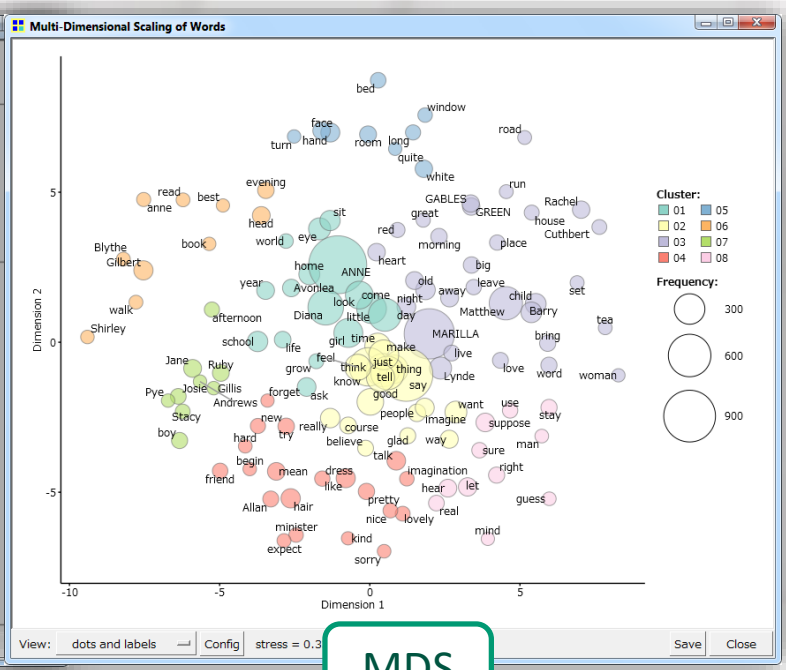
- To explore co-occurrences of words, you can also use:
  - ✓ hierarchical cluster analysis
  - ✓ Multi-dimensional scaling



co-occurrence network



cluster analysis



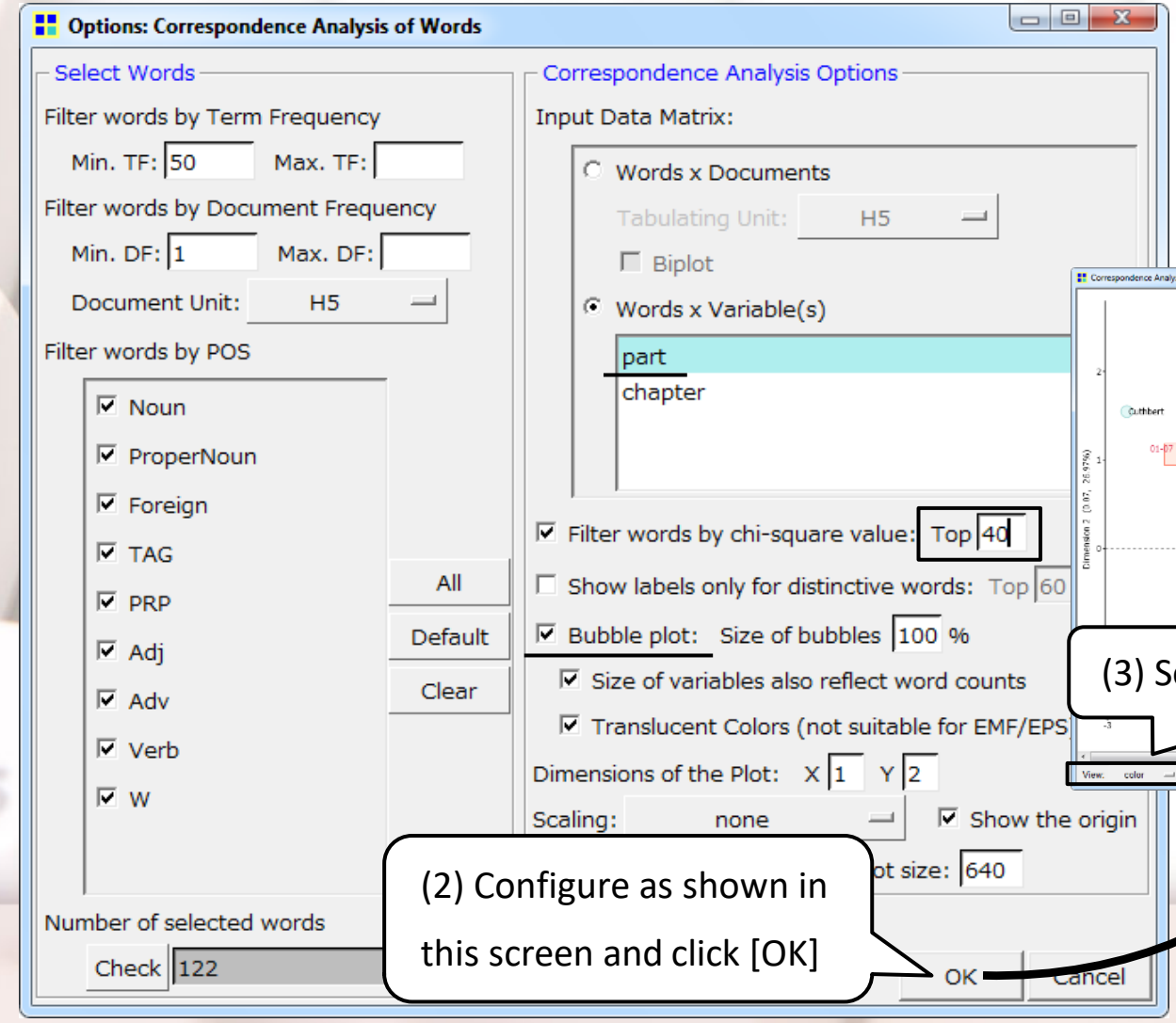
MDS

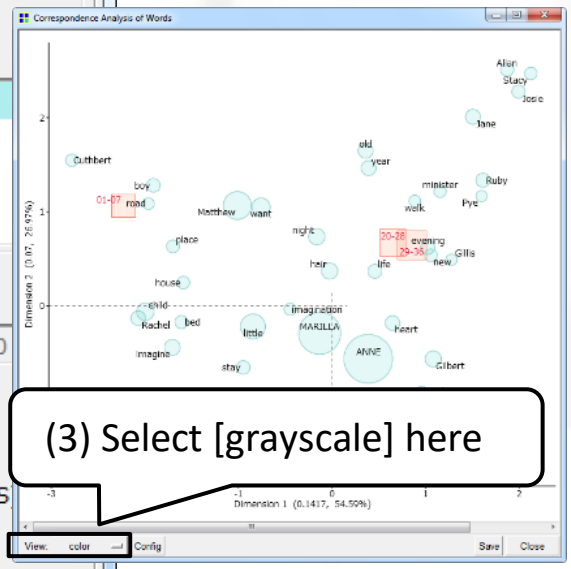
- By interpreting these result, you may find major themes of the text from groups of words which tend to appear together.
- KH Coder uses R as back end to execute these multivariate methods.

Step 1

Correspondence Analysis of Words (1/2)

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





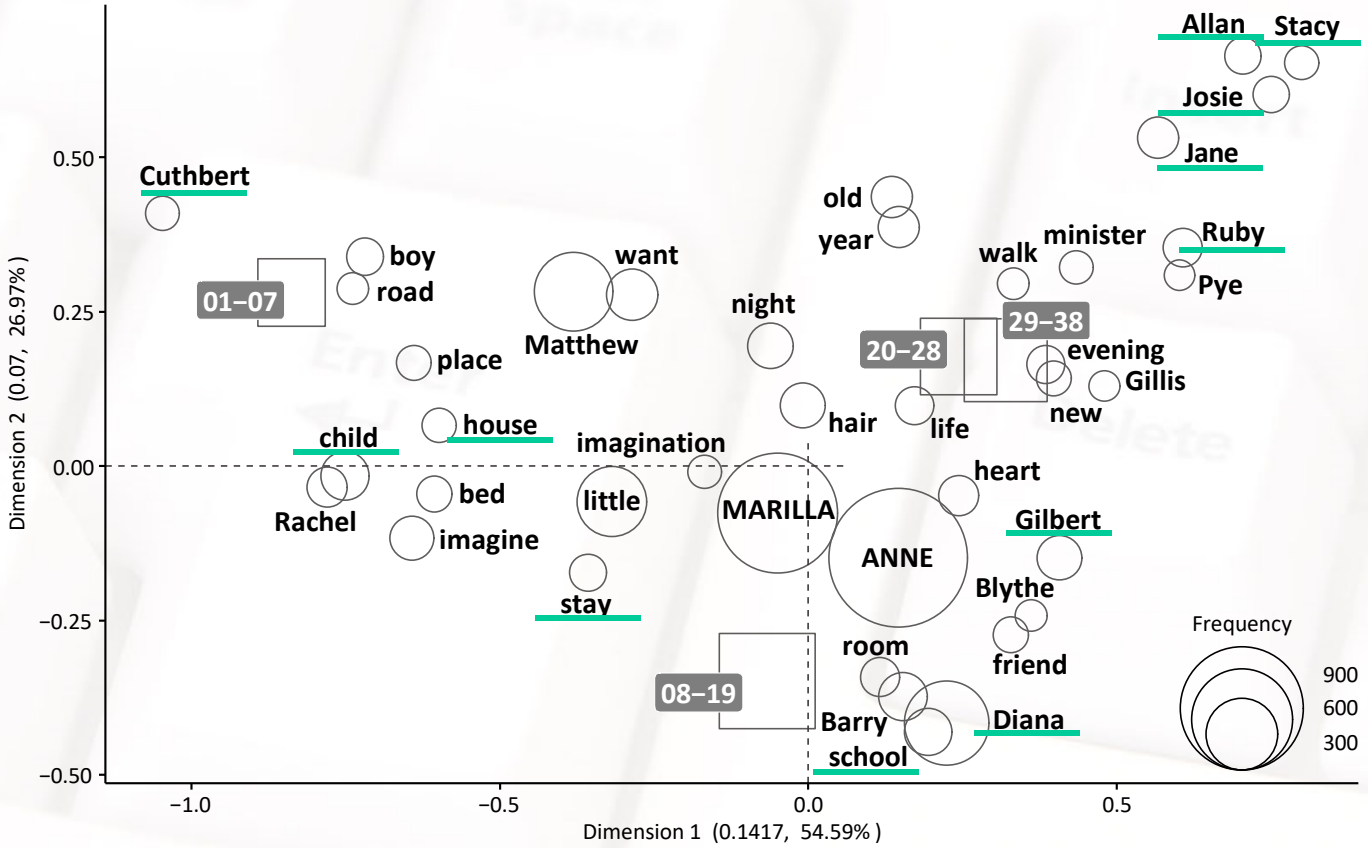
(3) Select [grayscale] here

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

Step 1

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”.

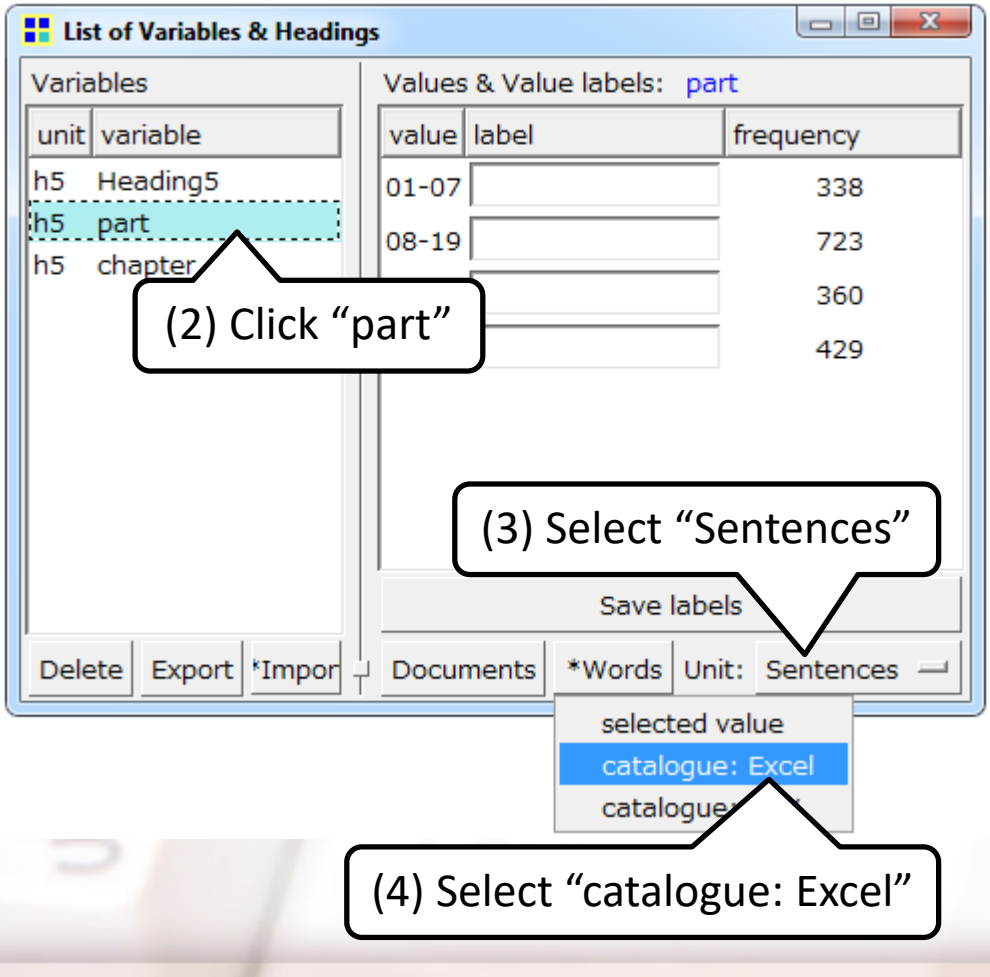


We can understand the story flow throughout the novel by checking characteristic words of each part.

Step 1

Characteristic Words of each Part

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



01-07		08-19		20-28	
say	.087	ANNE	.151	ANNE	.104
Matthew	.075	MARILLA	.114	MARILLA	.096
little	.045	Diana	.085	think	.061
come	.045	just	.053	make	.049
know	.042	little	.043	know	.048
child	.038	tell	.039	just	.048
thing	.036	school	.028	good	.039
look	.033	Barry	.027	Allan	.036
girl	.033	Lynde	.025	tell	.034
Spencer	.032	child	.022	thing	.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 “cording rules” like this:

Indicates the name of this code: “Character_name_Gilbert”

*Character_name_Gilbert
Gilbert or Gil

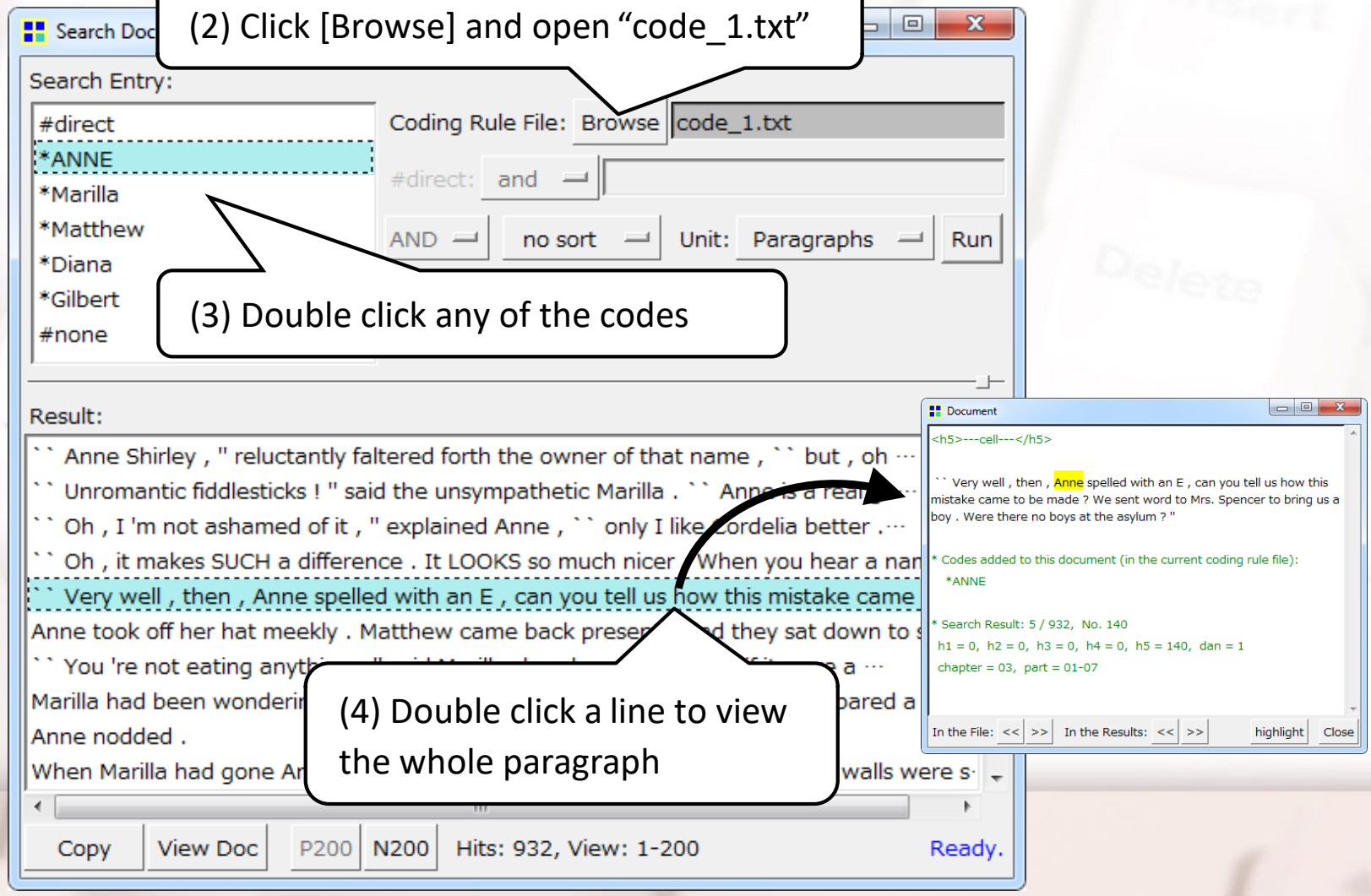
Not only the documents containing “Gilbert” but also those containing “Gil” are assigned this code.

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

Step 2

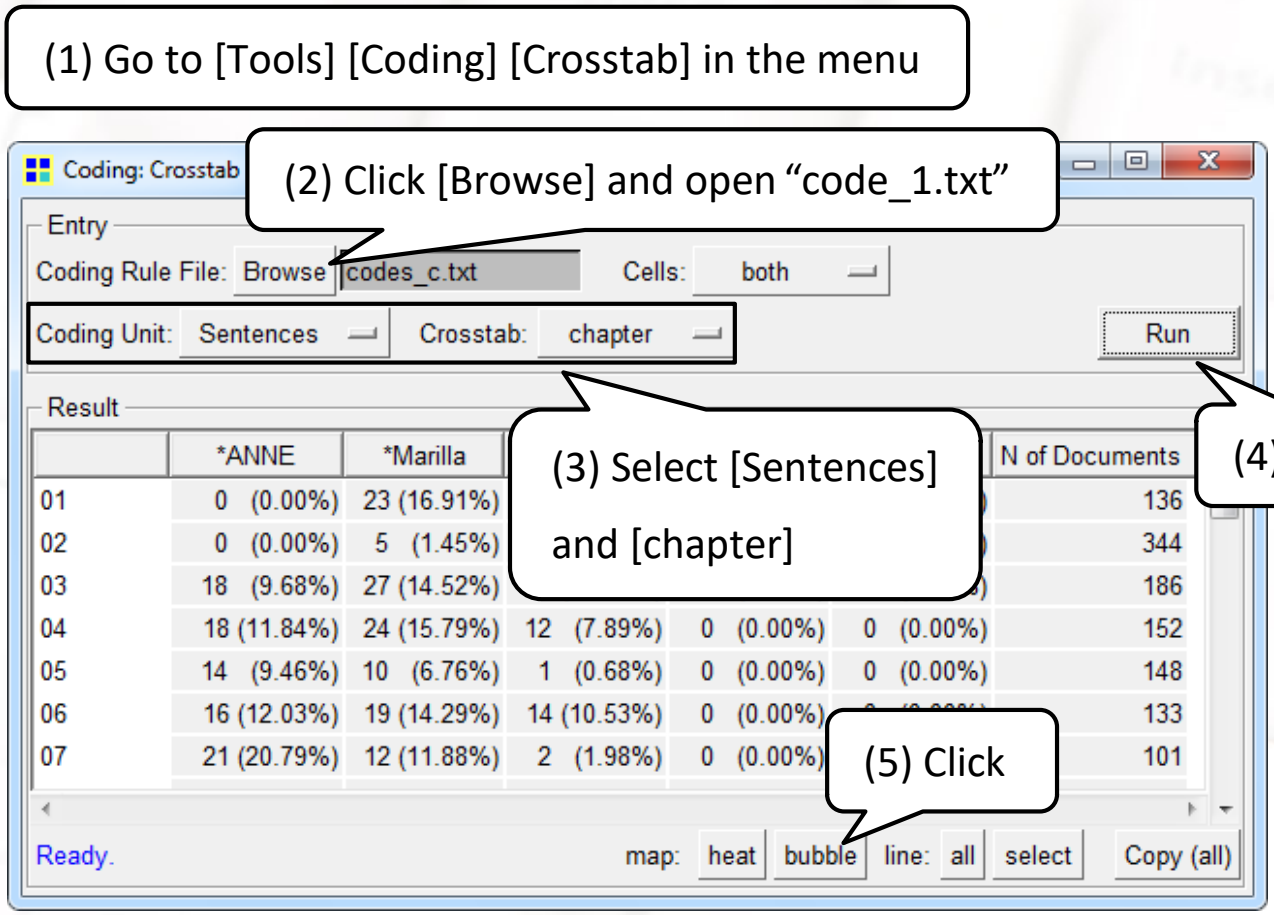
Retrieve Documents Assigned a Specific Code

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



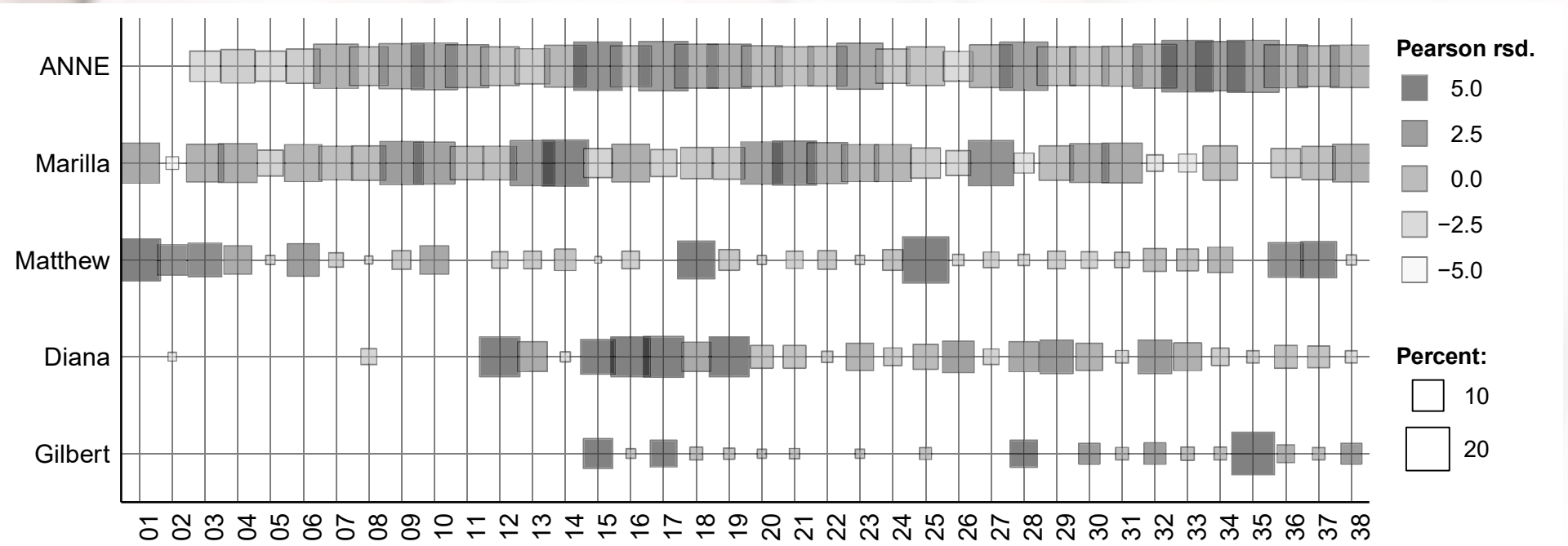
Characters in Each Chapter (1/2)

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



	*ANNE	*Marilla				N of Documents
01	0 (0.00%)	23 (16.91%)				136
02	0 (0.00%)	5 (1.45%)				344
03	18 (9.68%)	27 (14.52%)				186
04	18 (11.84%)	24 (15.79%)	12 (7.89%)	0 (0.00%)	0 (0.00%)	152
05	14 (9.46%)	10 (6.76%)	1 (0.68%)	0 (0.00%)	0 (0.00%)	148
06	16 (12.03%)	19 (14.29%)	14 (10.53%)	0 (0.00%)		133
07	21 (20.79%)	12 (11.88%)	2 (1.98%)	0 (0.00%)		101

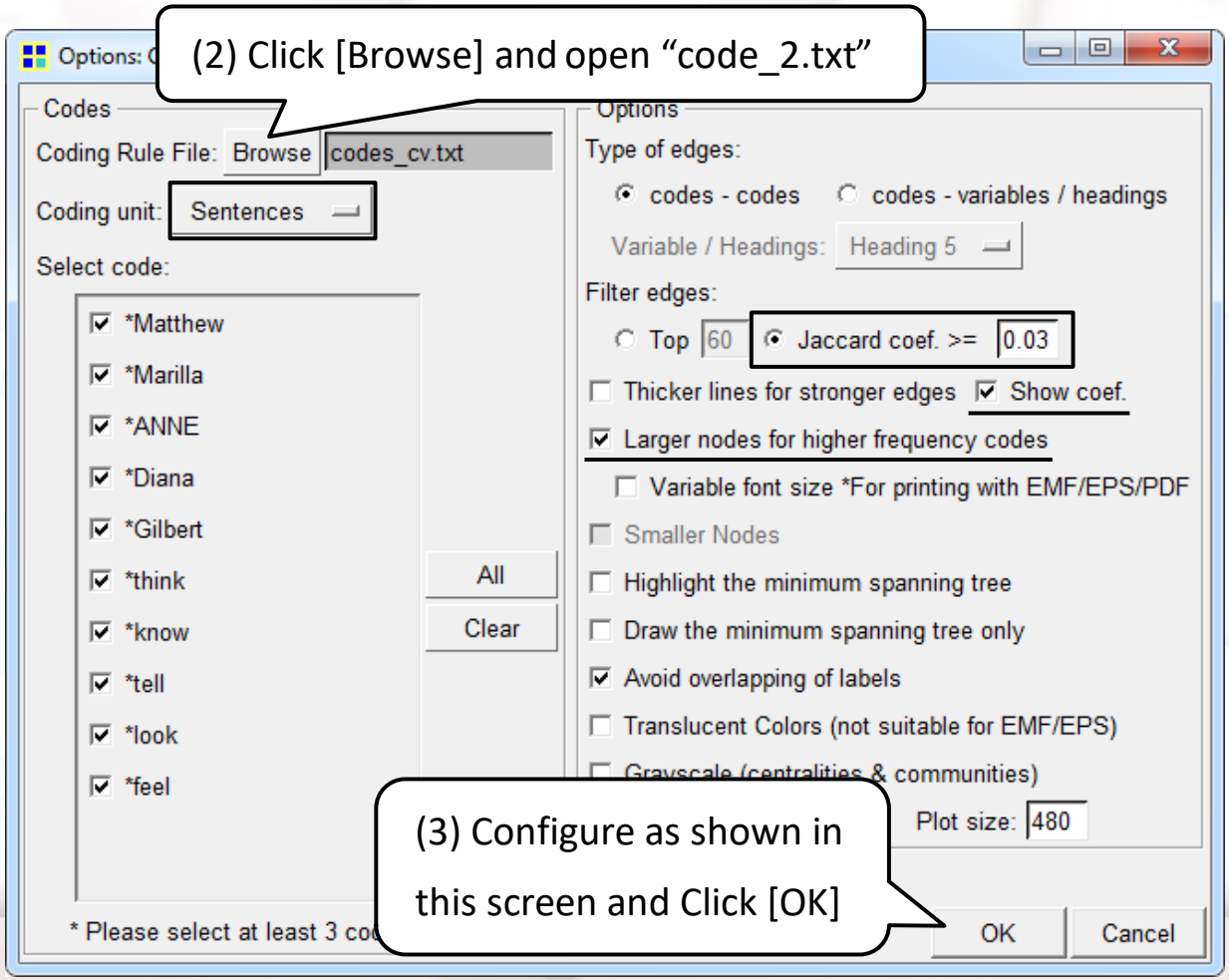
Characters in Each Chapter (2/2)



- 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!”

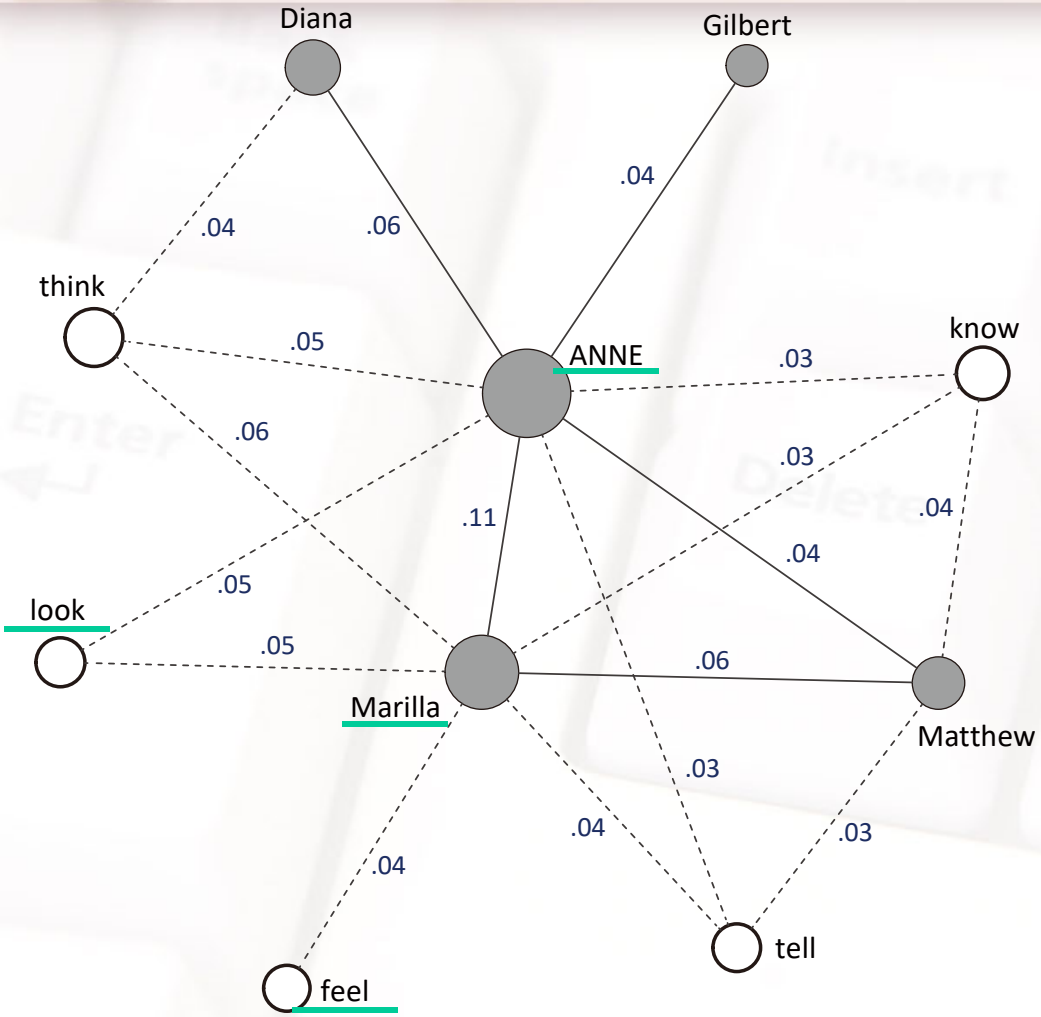
Characters and Verbs (1/2)

(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)

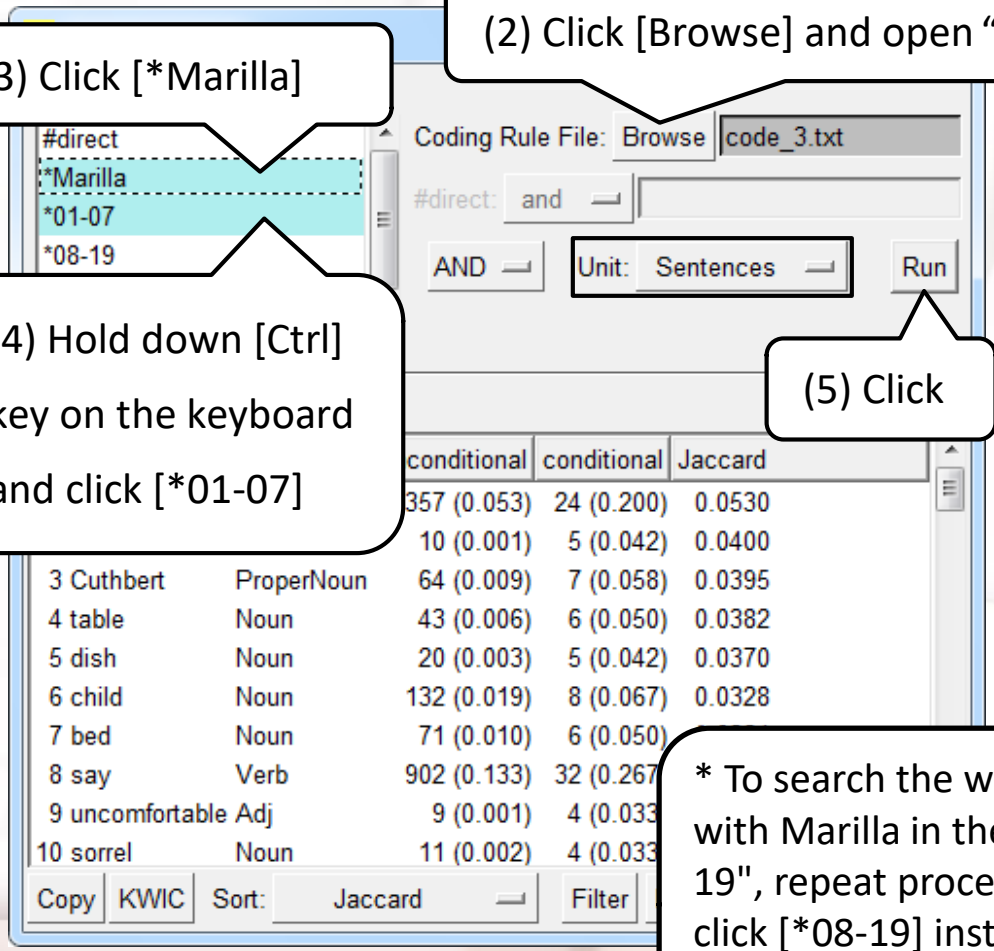
(1) Go to [Tools] [Words] [Word Association] in the menu

(2) Click [Browse] and open “code_3.txt”

(3) Click [*Marilla]

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

(5) Click



* 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).

Step 2

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	say	.072	say	.042	Matthew	.041
mare	.040	ANNE	.059	think	.034	look	.040
Cuthbert	.040	just	.039	ANNE	.032	sit	.039
table	.038	think	.036	cake	.030	ANNE	.038
dish	.037	brooch	.031	make	.028	say	.038
child	.033	tell	.030	minister	.028	face	.031
bed	.032	evening	.025	Allan	.026	girl	.026
say	.032	home	.024	feel	.025	think	.024
uncomfortable	.032	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/