워드스미스 툴 6.0 특강

A Special Seminar on WordSmith Tools 6.0

Dr. Mike Scott

University of Aston and Lexical Analysis Software Ltd.

mike.scott@aston.ac.uk

www.lexically.net

일 시: 2014. 8. 20(수) 16:00 ~ 17:30

장 소: 부산대학교 인문관 컴퓨터실(401호)

주 최: 코퍼스언어학연구회
(Research Association of Corpus Linguistics)

주 관: 부산대학교 글로벌 영상번역 창의인재 교육-연구 혁신사업단
(BK플러스)

Pusan National University

워드스미스 툴 6.0 특강

A Special Seminar on WordSmith Tools 6.0

Dr. Mike Scott

University of Aston and Lexical Analysis Software Ltd.

mike.scott@aston.ac.uk

www.lexically.net

일 시: 2014. 8. 20(수) 16:00 ~ 17:30

장 소: 부산대학교 인문관 컴퓨터실(401호)

주 최: 코퍼스언어학연구회
(Research Association of Corpus Linguistics)

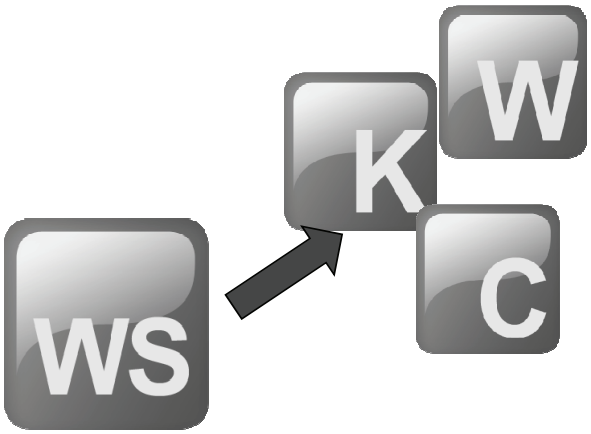
주 관: 부산대학교 글로벌 영상번역 창의인재 교육-연구 혁신사업단
(BK플러스)

Pusan National University

Contents

Developing WordSmith Tools

- Mike Scott
- Aston University
- Lexical Analysis Software Ltd



1

For WordSmith you may if you wish use the single user licence for the individual PCs, but it may be easier to use a network licence:

Here is the network registration code.

Registered to: Korea Maritime and Ocean University
Other details: temporary
Registration code: SD00.0617.2331.6428.1570.6688.A8AU

WordSmith version: 06
For site use, not individual use
(http://lexically.net/downloads/version6/HTML/index.html?networkdefaults.htm)
see
http://www.lexically.net/wordsmith/version6/faqs/network_installation.htm for detailed instructions on installation.

Issued: 13 August 2014
This registration OK until: 14 September 2014
Download from http://lexically.net/wordsmith/version6/

2

The aim

- personal use at home
- many languages
- assortment of tools
- lexical focus
- download only

3

History

- *MicroConcord* 1993 (with Tim Johns)
- *WordSmith Tools* 1.0 1996, Oxford University Press
  - 2.0, 1997 OUP
  - 3.0, 1999 OUP
- 4.0, 2004 OUP (Unicode)
  - 5.0, 2008 Lexical Analysis Software (WSConcgram)
  - 6.0, 2012 Lexical Analysis Software (clouds, scripts, phrase frames, colour categories)

4

Learning

- Key words function developed almost by accident
- Many settings & features arose at user request

5

Problems

- developing an intuitive interface
- providing help
- ignoring language-specific knowledge
- technology constantly changing

6

Demonstration and hands-on

- See http://www.lexically.net/wordsmith/support/get_started_guides.html if you need help...

- a word list
- a concordance from the word list
- viewing the source text
- a key word list
- exporting results to Word or Excel

7

Help

- http://www.lexically.net/wordsmith/support/get_started_guides.html
- (no Guide yet in Korean)
- http://www.lexically.net/downloads/version6/HTML/index.html
- free licence:

```
Registered to: Korea Maritime and Ocean University
Other details: temporary
Registration code: SA00.0613.2777.1799.1570.8653.A8A3
```

This registration is OK until: 14 September 2014

Download from http://lexically.net/wordsmith/version6/

When registering (Utilities | Registration), please PASTE these details in EXACTLY as you see them here.

8

How to Use WordSmith Tools:

Install WS, Wordlist, Keywords, Concord, Index, Clusters, Keyclusters

August 20, 2014

Se-Eun Jhang & Sungmin Lee
(Korea Maritime and Ocean University)

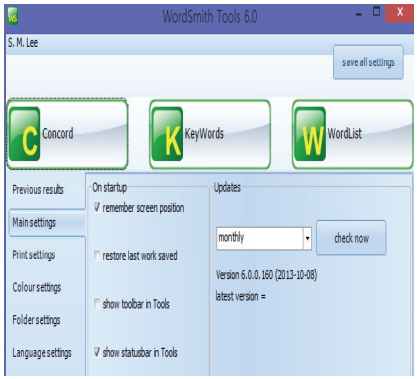
Busan National University

Contents

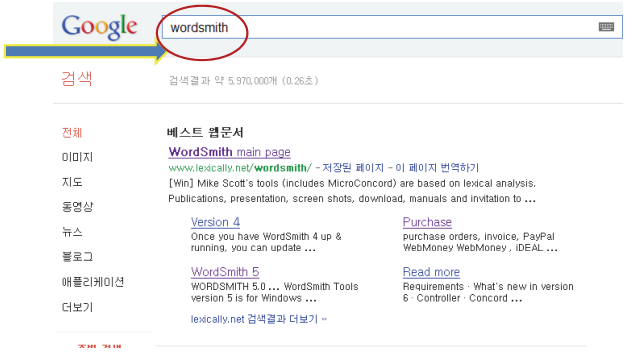
- How to install WordSmith
- How to get Wordlist
- How to get Keywords
- How to get Concordance
- How to get Index
- How to get Clusters
- How to get Keyclusters

Advantages of Wordsmith Tools

- 편리한 화면 구성
- 상세한 매뉴얼
- 다양한 통계제공



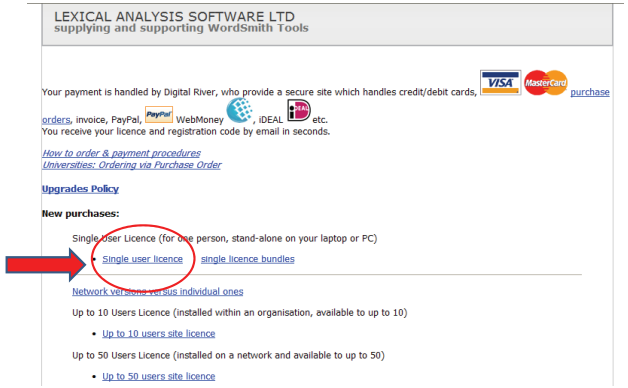
www.lexically.net/wordsmith



Purchase WS

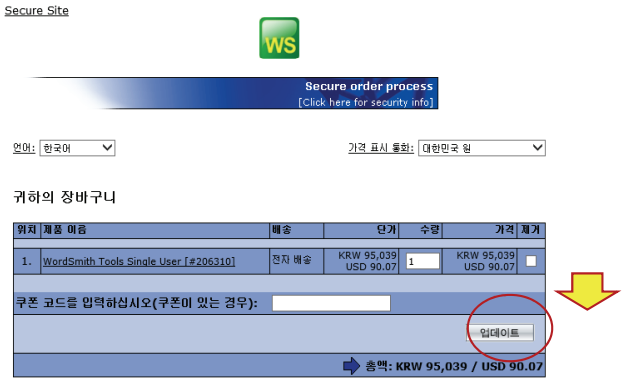
5

Purchase WS



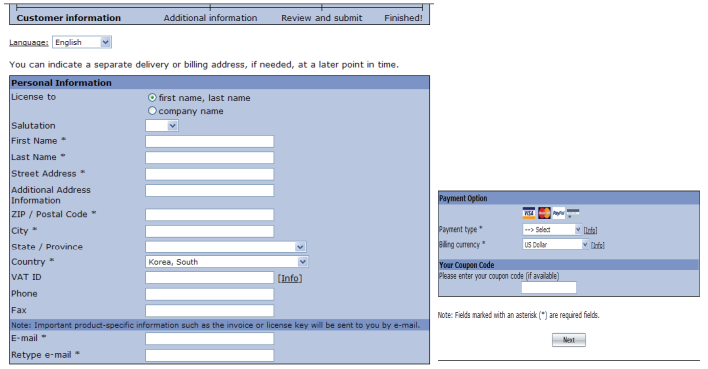
6

Single User License



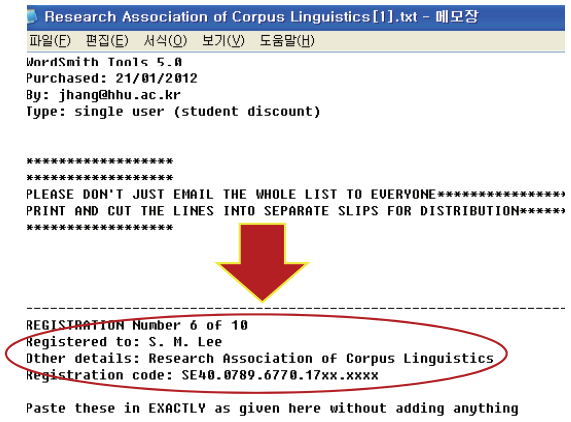
7

Single User License



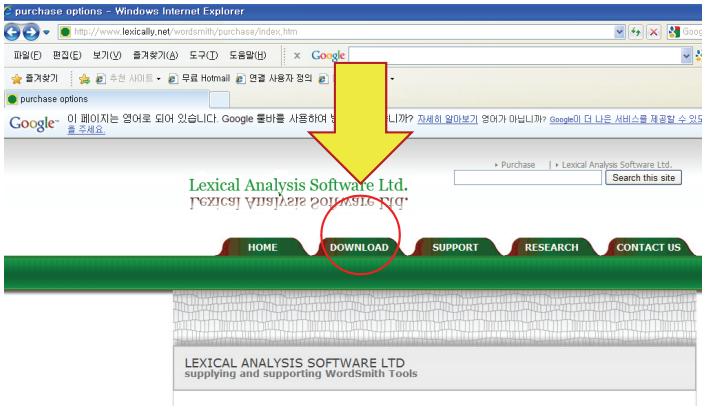
8

License Key



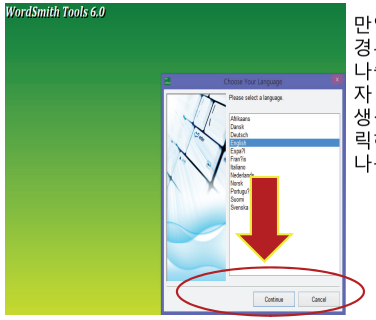
9

Installation of WS

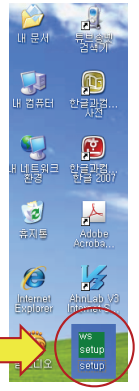


10

Installation of WS 6.0

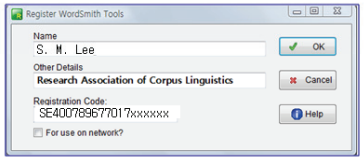


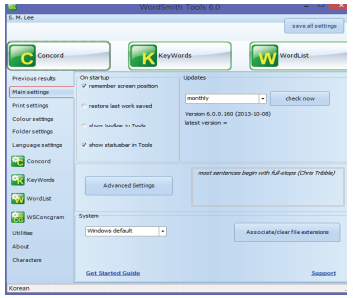
만약 계속하지 못하는 경우가 생기면 Cancel 나중에 계속 설치하고자 하면 바탕화면에 생성된 WS setup 클릭하면 좌측화면이 또 나옴



11

Installation of WS





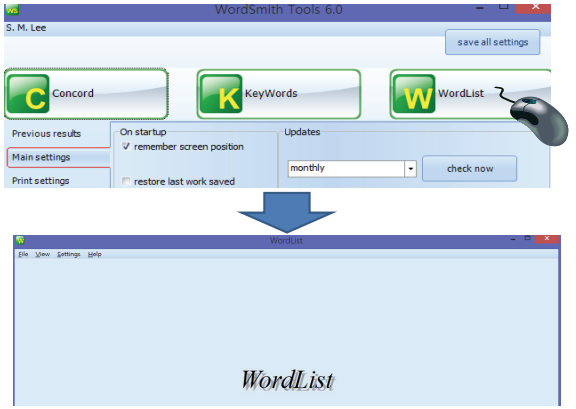
12

Wordlist

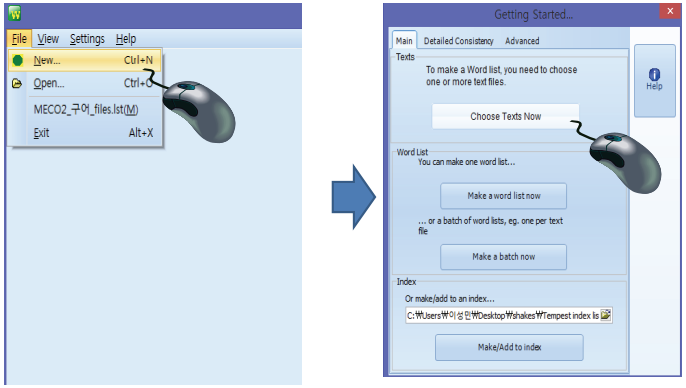
- This program generates word lists based on one or more plain text files.
- Word lists are shown both in alphabetical and frequency order.
- They can be saved for later use, edited, printed, copied to your word-processor, or saved as text files.

출처 : Wordsmith 6.0 manual

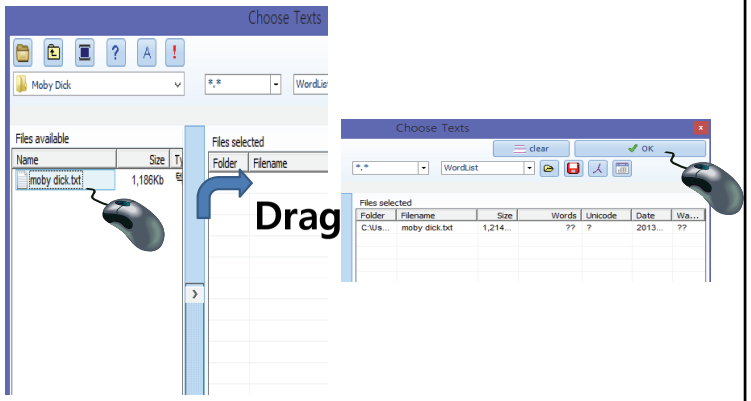
WordList



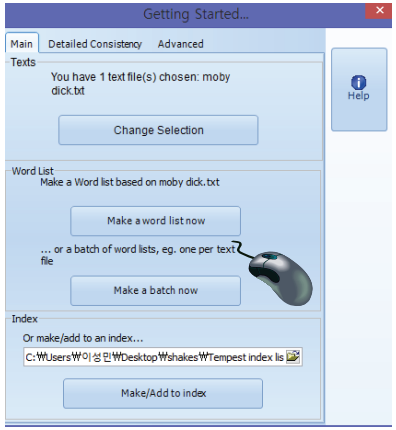
Start WordList



Upload Txt files

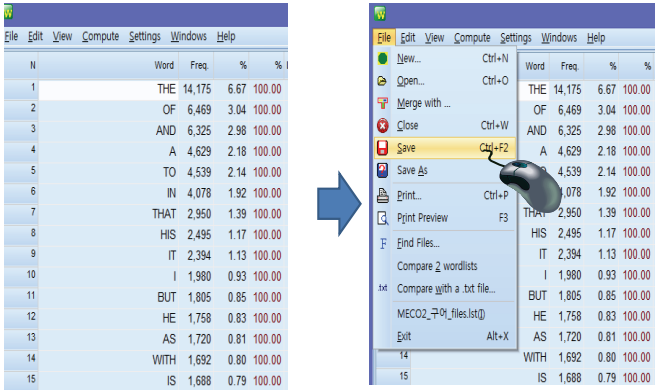


Click "Make a wordlist now"



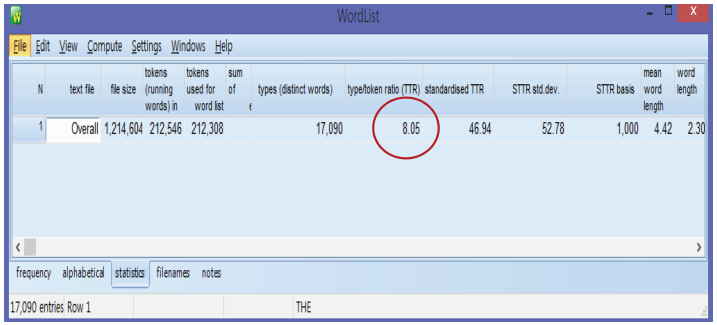
17

Save WordList



18

Check Statistics for Moby Dick



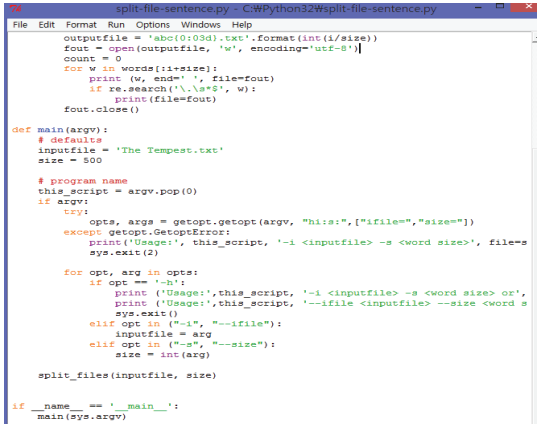
19

TTR (Type-Token Ratio)

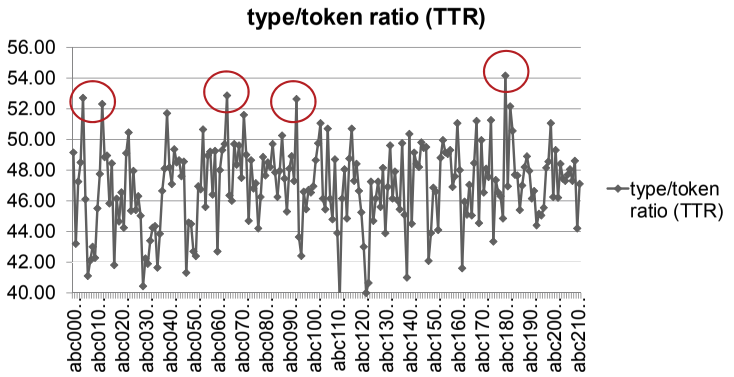
- TTR = (number of type / number of token) x 100
- TTR means the ratio of type and token. The way to get the TTR is the number of types over the number of tokens multiplied by 100.
- For example, the TTR of texts used by native speakers is 22.67. On the other hand, the TTR of texts by non-native speakers is 13.48. This shows that native speakers use a greater variety of vocabulary than non-native speakers.

20

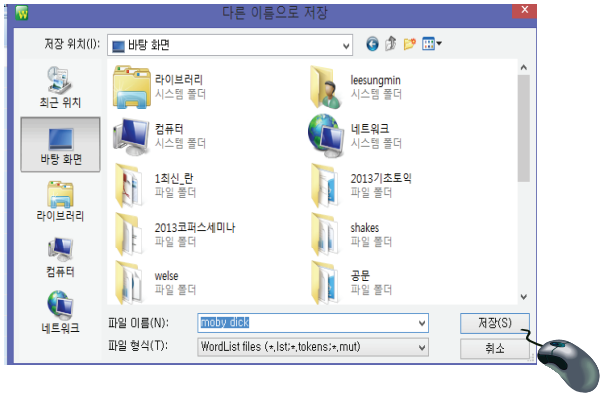
Python Coding For Customized TTR Analysis



Moby Dick TTR ratio per 100 words



Save a Wordlist Files



Keyness

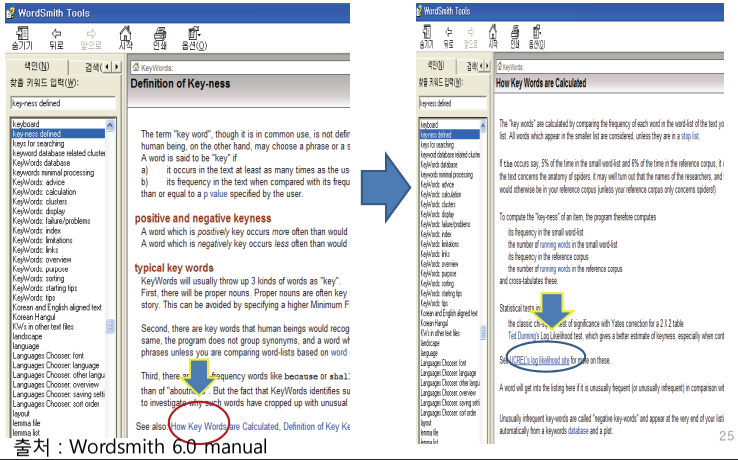
The term "key word", though it is in common use, is not defined in Linguistics. This program identifies key words on a mechanical basis by comparing patterns of frequency. (A human being, on the other hand, may choose a phrase or a superordinate as a key word.) A word is said to be "key" if

a) it occurs in the text at least as many times as the user has specified as a Minimum Frequency
b) its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p value specified by the user.

positive and negative keyness
A word which is *positively* key occurs *more* often than would be expected by chance in comparison with the reference corpus.
A word which is *negatively* key occurs *less* often than would be expected by chance in comparison with the reference corpus.

출처 : Wordsmith 6.0 manual

How Are Keywords Calculated?

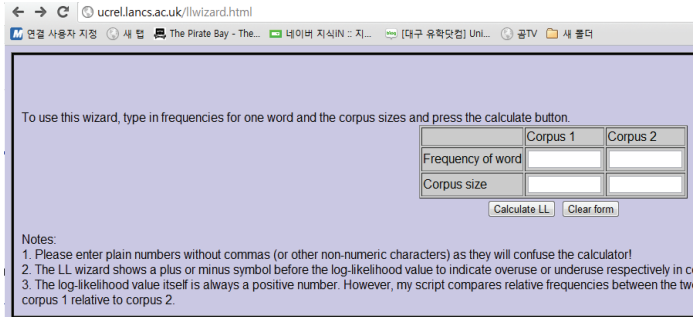


출처 : Wordsmith 6.0 manual

25

Log-likelihood Calculator

http://ucrel.lancs.ac.uk/llwizard.html



26

How to Calculate Log Likelihood

Log likelihood is calculated by constructing a contingency table as follows:

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

Note that the value 'c' corresponds to the number of words in corpus one, and 'd' corresponds to the number of words in c
observed values (O), whereas we need to calculate the expected values (E) according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In our case N1 = c, and N2 = d. So, for this word, E1 = c*(a+b) / (c+d) and E2 = d*(a+b) / (c+d). The calculation for the exp
so we do not need to normalize the figures before applying the formula. We can then calculate the log-likelihood value ac

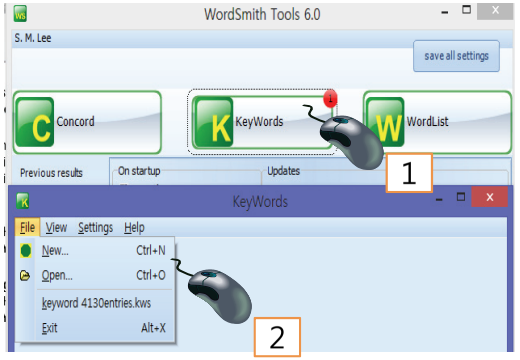
$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

The higher the G2 value, the more significant is the difference between two frequency scores. For these tables, a
G2 of 3.8 or higher is significant at the level of p < 0.05 and a G2 of 6.6 or higher is significant at p < 0.01.

- 95th percentile; 5% level; p < 0.05; critical value = 3.84
- 99th percentile; 1% level; p < 0.01; critical value = 6.63
- 99.9th percentile; 0.1% level; p < 0.001; critical value = 10.83
- 99.99th percentile; 0.01% level; p < 0.0001; critical value = 15.13

27

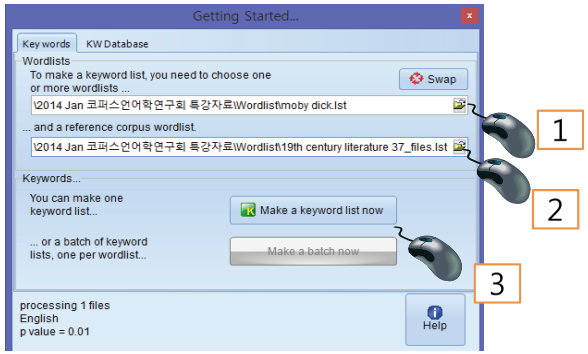
Getting Keywords



28

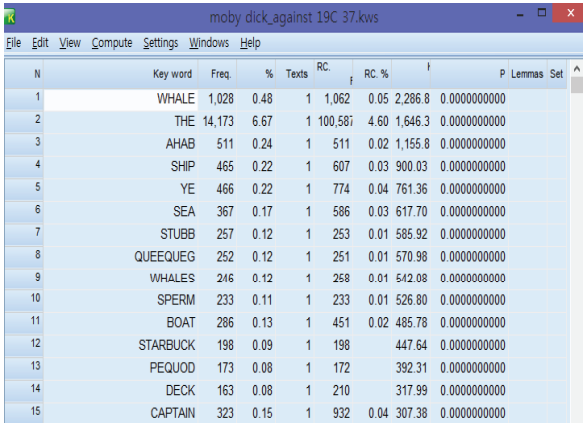
Moby Dick vs Contem19C

Click Target Corpus and Reference Corpus
Click Make a keyword list now



29

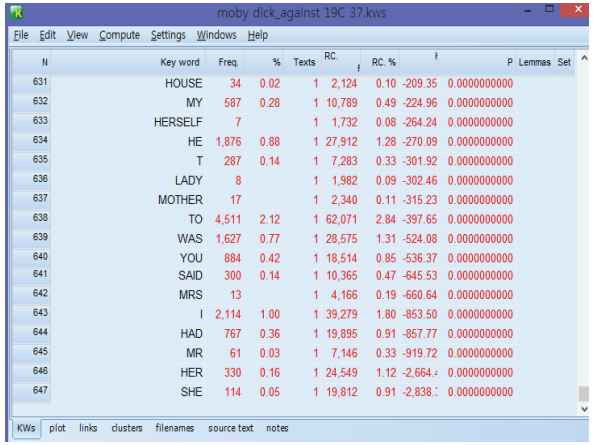
Positive Keywords



N	Key word	Freq.	%	Texts	RC.	RC. %		P	Lemmas	Set
1	WHALE	1,028	0.48	1	1,062	0.05	2,286.8	0.0000000000		
2	THE	14,173	6.67	1	100,587	4.60	1,646.3	0.0000000000		
3	AHAB	511	0.24	1	511	0.02	1,155.8	0.0000000000		
4	SHIP	465	0.22	1	607	0.03	900.03	0.0000000000		
5	YE	466	0.22	1	774	0.04	761.36	0.0000000000		
6	SEA	367	0.17	1	586	0.03	617.70	0.0000000000		
7	STUBB	257	0.12	1	253	0.01	585.92	0.0000000000		
8	QUEEQUEG	252	0.12	1	251	0.01	570.98	0.0000000000		
9	WHALES	246	0.12	1	258	0.01	542.08	0.0000000000		
10	SPERM	233	0.11	1	233	0.01	526.80	0.0000000000		
11	BOAT	286	0.13	1	451	0.02	485.78	0.0000000000		
12	STARBUCK	198	0.09	1	198		447.64	0.0000000000		
13	PEQUOD	173	0.08	1	172		392.31	0.0000000000		
14	DECK	163	0.08	1	210		317.99	0.0000000000		
15	CAPTAIN	323	0.15	1	932	0.04	307.38	0.0000000000		

30

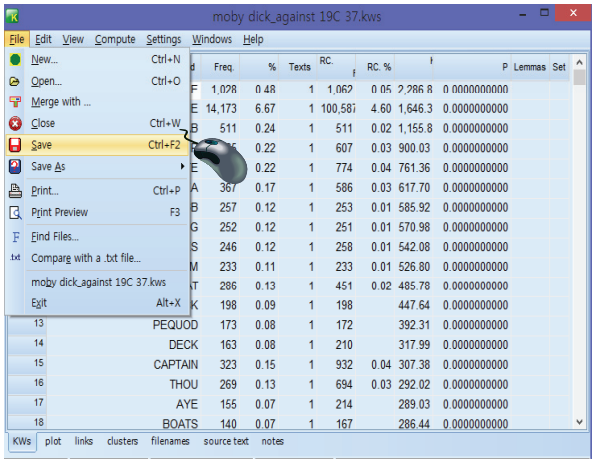
Negative Keywords



N	Key word	Freq.	%	Texts	RC.	RC. %		P	Lemmas	Set
631	HOUSE	34	0.02	1	2,124	0.10	-209.35	0.0000000000		
632	MY	587	0.28	1	10,789	0.49	-224.96	0.0000000000		
633	HERSELF	7		1	1,732	0.08	-264.24	0.0000000000		
634	HE	1,876	0.88	1	27,912	1.28	-270.09	0.0000000000		
635	T	287	0.14	1	7,283	0.33	-301.92	0.0000000000		
636	LADY	8		1	1,982	0.09	-302.46	0.0000000000		
637	MOTHER	17		1	2,340	0.11	-315.23	0.0000000000		
638	TO	4,511	2.12	1	62,071	2.84	-397.65	0.0000000000		
639	WAS	1,627	0.77	1	28,575	1.31	-524.08	0.0000000000		
640	YOU	884	0.42	1	18,514	0.85	-536.37	0.0000000000		
641	SAID	300	0.14	1	10,365	0.47	-645.53	0.0000000000		
642	MRS	13		1	4,166	0.19	-660.64	0.0000000000		
643	I	2,114	1.00	1	39,279	1.80	-853.50	0.0000000000		
644	HAD	767	0.36	1	19,895	0.91	-857.77	0.0000000000		
645	MR	61	0.03	1	7,146	0.33	-919.72	0.0000000000		
646	HER	330	0.16	1	24,549	1.12	-2,664.4	0.0000000000		
647	SHE	114	0.05	1	19,812	0.91	-2,838.3	0.0000000000		

31

How to Save as Keyword List



32

Concordance

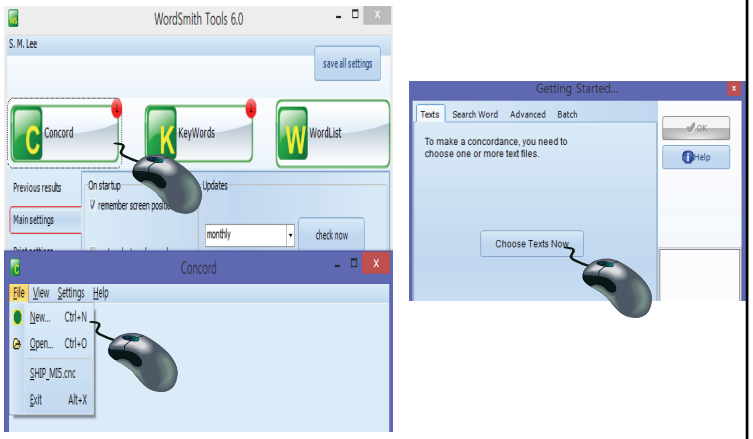
• To use it you will specify a search word, which Concord will seek in all the text files you have chosen.

• It will then present a concordance display, and give you access to information about collocates of the search word.

출처 : Wordsmith 6.0 manual

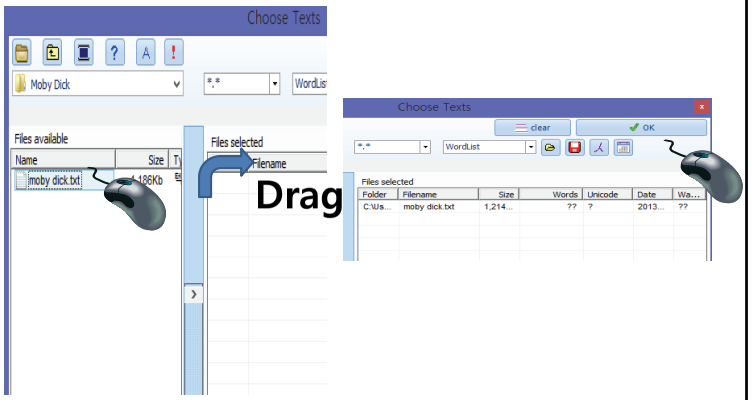
33

Concordance



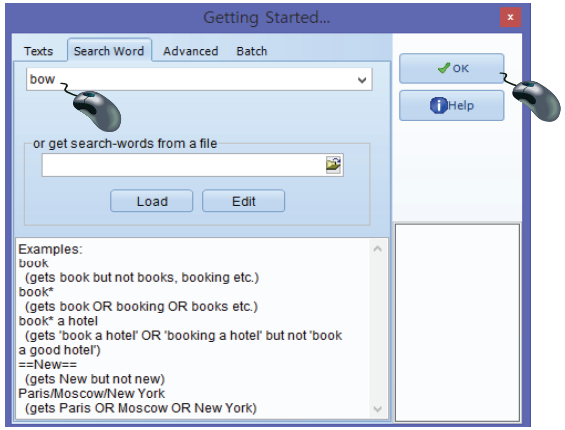
34

Upload Txt file



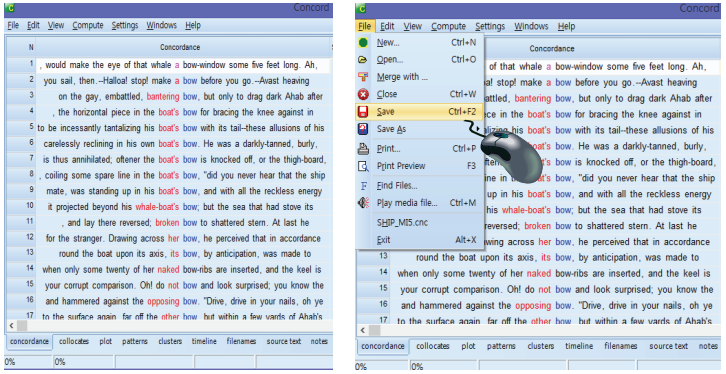
35

Insert Search Word

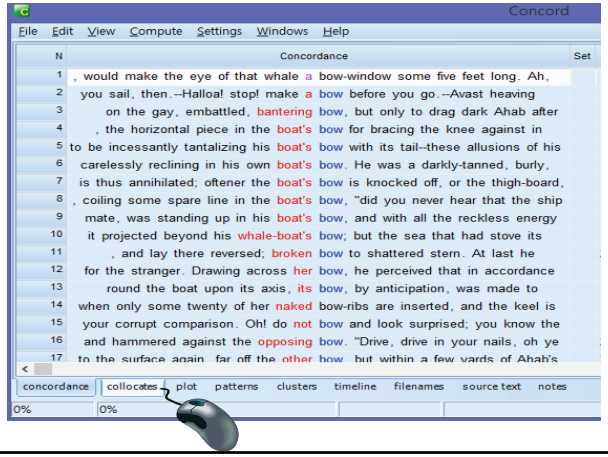


36

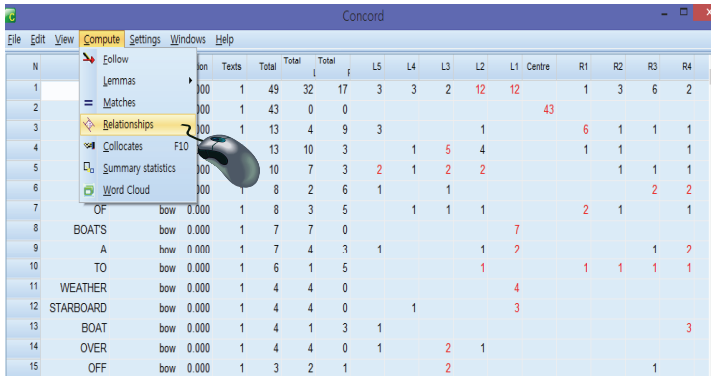
Save Concordance



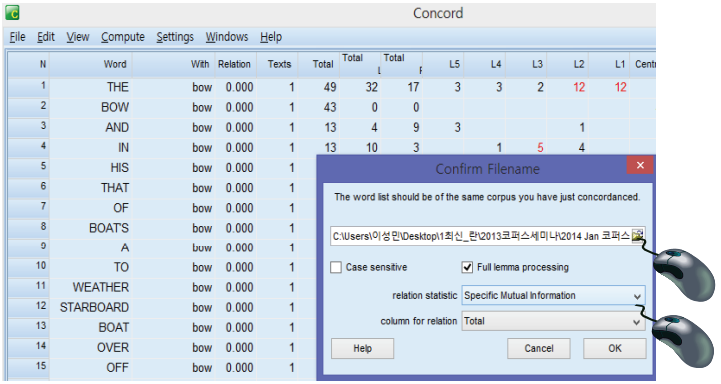
How to get Collocations



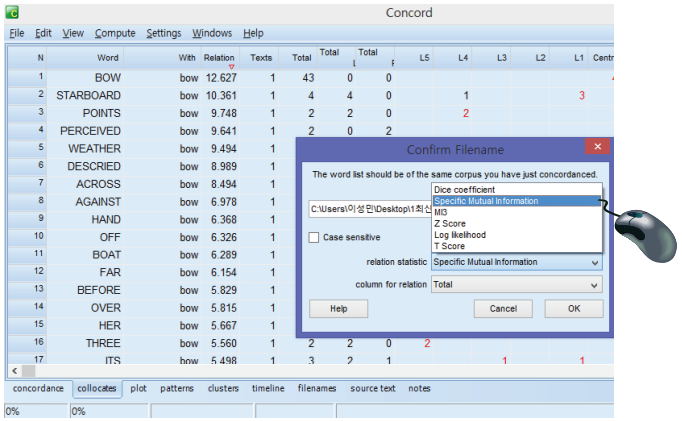
Click Compute - Relationships



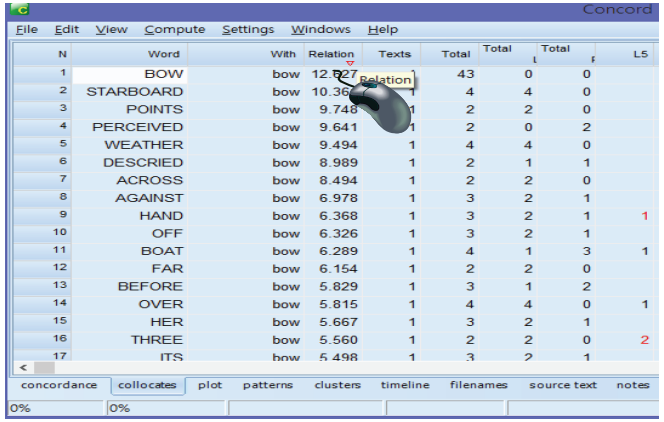
Upload Wordlist and Click Statistics



Kinds of Collocations

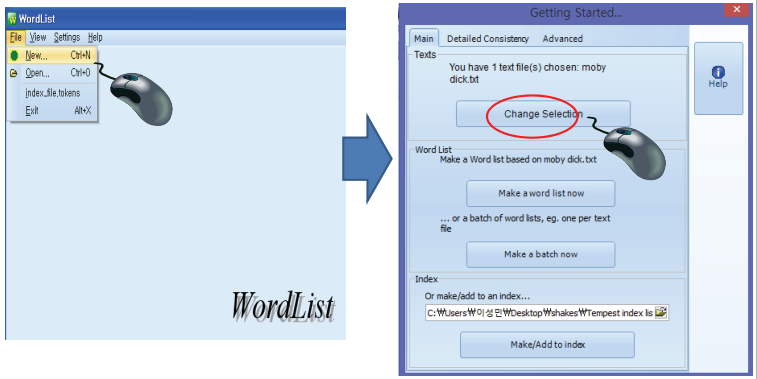


Sort Relation values

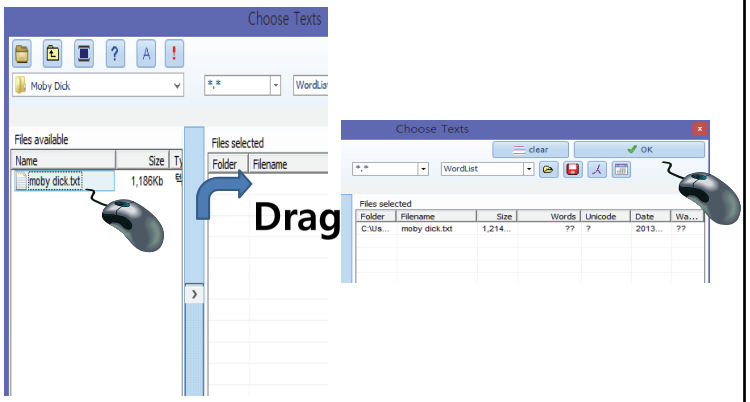


How to Create Index, Clusters, Keyclusters

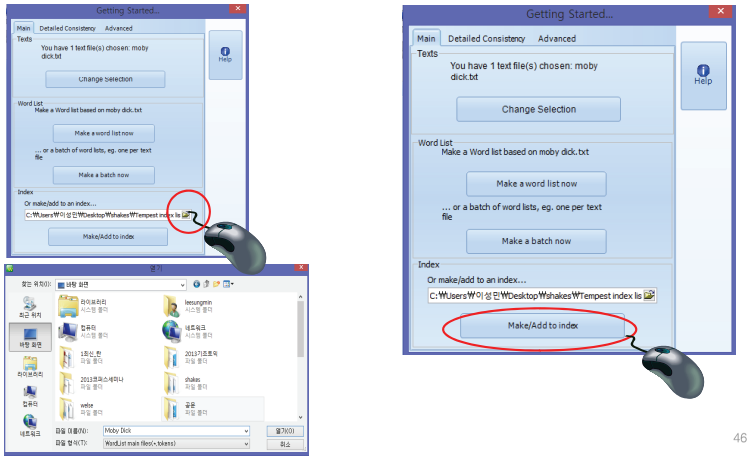
INDEX: Make a Wordlist tool index



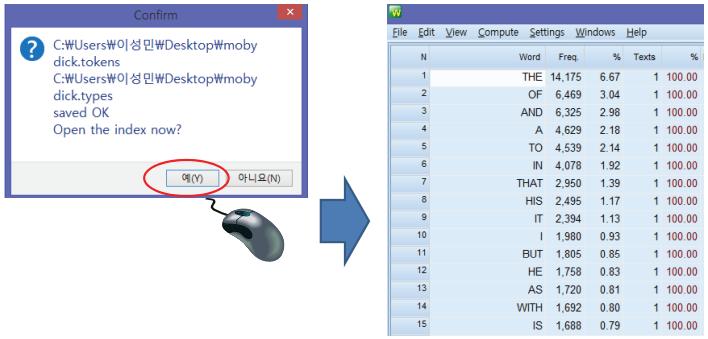
Upload Txt file



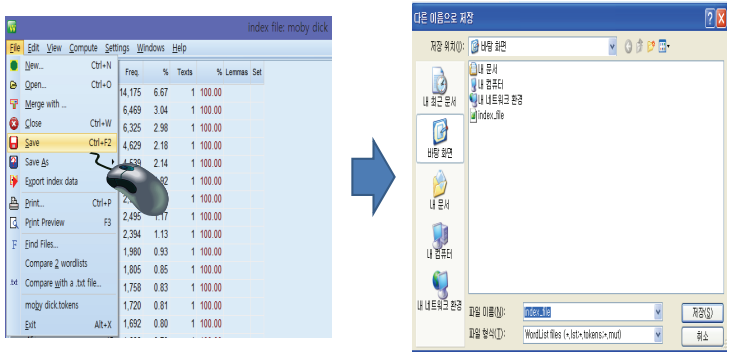
Save/Click Make/Add to Index



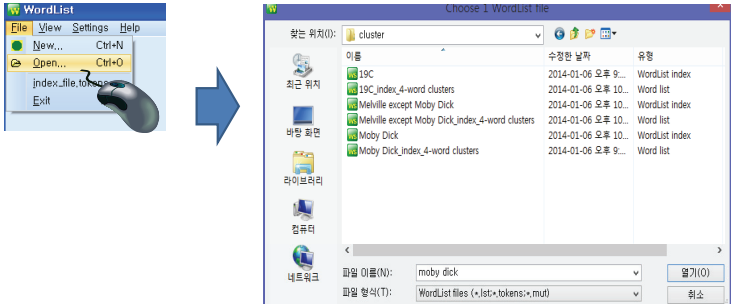
Create Index File



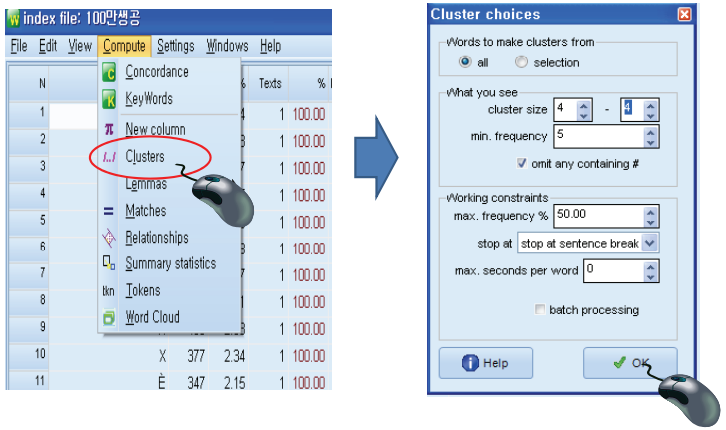
Save index files in your folder



Open saved index files



Go compute & click Cluster choices

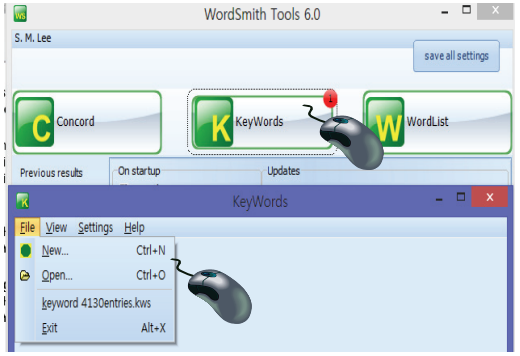


Output of clusters

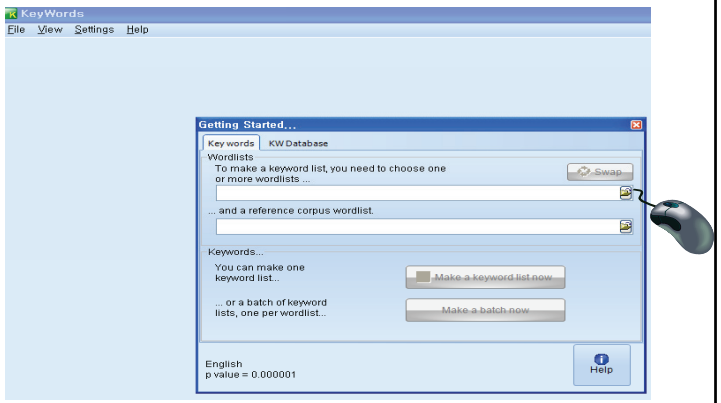
File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
1	OF THE SPERM WHALE	42	0.02	1	100.00
2	THE SPERM WHALE S	27	0.01	1	100.00
3	AT THE SAME TIME	20		1	100.00
4	OF THE WHALE S	18		1	100.00
5	THE BOTTOM OF THE	17		1	100.00
6	THE OLD MAN S	16		1	100.00
7	IN THE ACT OF	12		1	100.00
8	THE WHITE WHALE S	12		1	100.00
9	AND AT THE SAME	11		1	100.00
10	AS IF IT WERE	11		1	100.00
11	I DON T KNOW	11		1	100.00
12	IN THE CASE OF	11		1	100.00
13	OF THE WHITE WHALE	11		1	100.00
14	SPERM WHALE S HEAD	11		1	100.00
15	BE IT SAID THAT	10		1	100.00

Click a Keywords tool for making a key clusters list

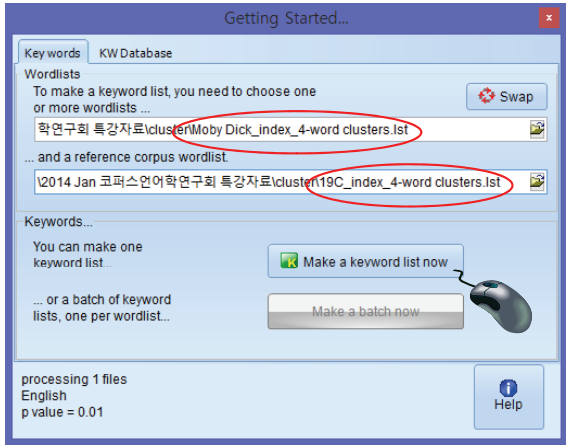


Start making Keyclusters



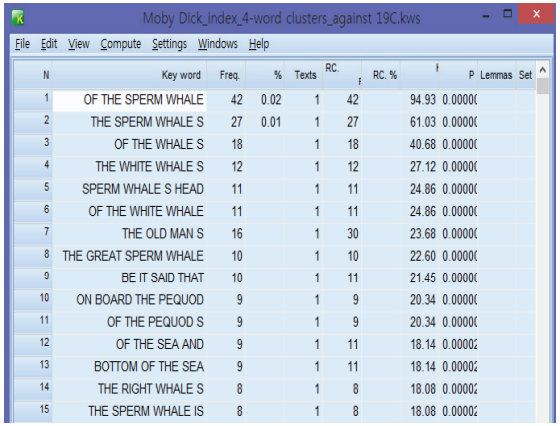
53

Upload clusters list



54

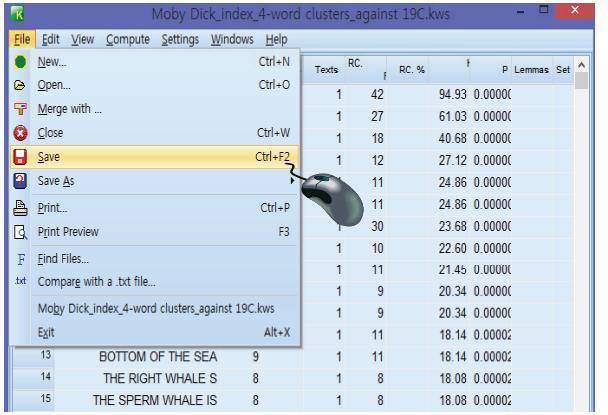
Output of Plus Keyclusters



N	Key word	Freq.	%	Texts	RC.	RC. %	P	Lemmas	Set
1	OF THE SPERM WHALE	42	0.02	1	42	94.93	0.0000(		
2	THE SPERM WHALE S	27	0.01	1	27	61.03	0.0000(		
3	OF THE WHALE S	18		1	18	40.68	0.0000(		
4	THE WHITE WHALE S	12		1	12	27.12	0.0000(		
5	SPERM WHALE S HEAD	11		1	11	24.86	0.0000(		
6	OF THE WHITE WHALE	11		1	11	24.86	0.0000(		
7	THE OLD MAN S	16		1	30	23.68	0.0000(		
8	THE GREAT SPERM WHALE	10		1	10	22.60	0.0000(		
9	BE IT SAID THAT	10		1	11	21.45	0.0000(		
10	ON BOARD THE PEQUOD	9		1	9	20.34	0.0000(		
11	OF THE PEQUOD S	9		1	9	20.34	0.0000(		
12	OF THE SEA AND	9		1	11	18.14	0.0000(		
13	BOTTOM OF THE SEA	9		1	11	18.14	0.0000(		
14	THE RIGHT WHALE S	8		1	8	18.08	0.0000(		
15	THE SPERM WHALE IS	8		1	8	18.08	0.0000(		

55

How to Save Keyclusters List



56