

# **Step-by-step guide to WordSmith Tools**

## **Translation: CK Jung**

[https://lexically.net/wordsmith/step\\_by\\_step\\_Korean7/index.html](https://lexically.net/wordsmith/step_by_step_Korean7/index.html)

---

정재관 **Chae Kwan Jung** BEng(Hons) Birmingham MSc Warwick EdD Warwick Cert Oxford

Institute for Corpus Research, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon,  
22012, South Korea

E: [help@icr.or.kr](mailto:help@icr.or.kr) / T: +82 (0)32 835 8129 / H: <http://www.icr.or.kr>

---

## 1. 소개

이 튜토리얼은 워드스미스 툴스(이하, '워드스미스') 사용자의 편의를 위해 만들었고, 워드스미스의 사용자(이하, '사용자')의 이해를 돋기 위해 가급적 다양한 스크린 캡처를 포함하였다. <그림 1>은 워드스미스를 시작하면 나타나는 첫 화면이다.

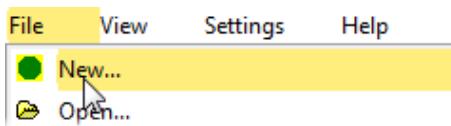


<그림 1> 워드스미스

<그림 1>에서 볼 수 있듯이, 워드스미스에는 위쪽에 3 개의 큰 버튼이 있고, 왼쪽에 사용자가 기타 설정을 조정할 수 있는 다양한 탭이 있다. 위쪽에 있는 3 개의 큰 버튼 중 콘코드(Concord)를 클릭하면 콘코던스를 할 수 있고, 키워드(KeyWord)를 클릭하면 코퍼스에서 키워드를 추출할 수 있고, 워드리스트(WordList)를 클릭하면 코퍼스에서 워드 리스트를 추출할 수 있다. 워드스미스를 시작하기 위해서는 이 3 개의 큰 버튼 중 하나를 선택한다.

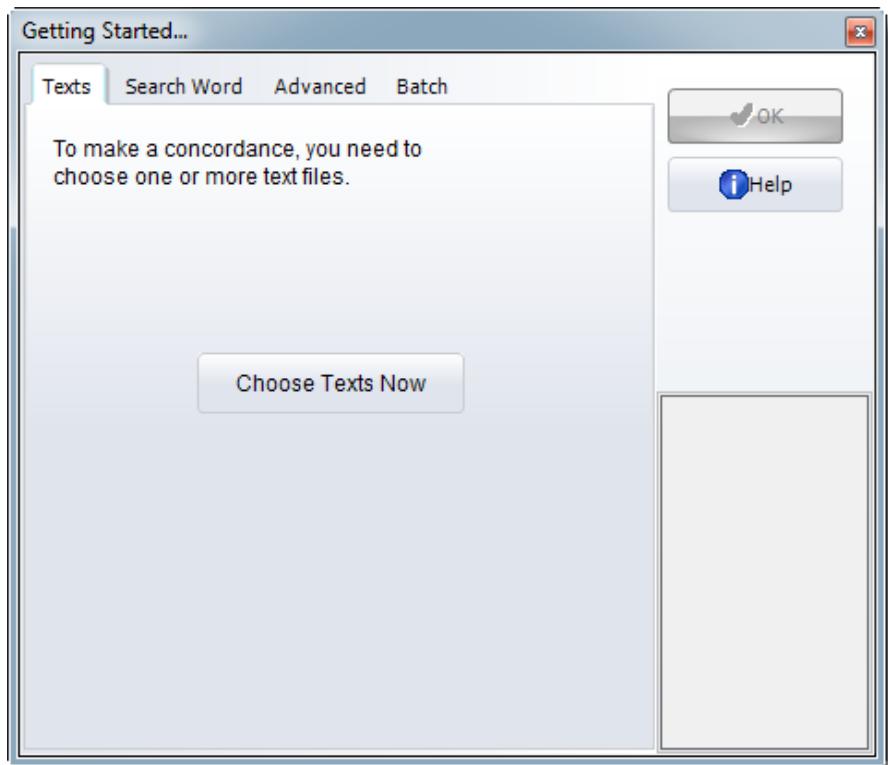
## 2. 텍스트 선택하기

워드스미스를 시작한 다음 사용자가 분석하고자 하는 텍스트를 불러와 보자. 우선 메뉴에서 *File / New*를 선택한다(<그림 2>).



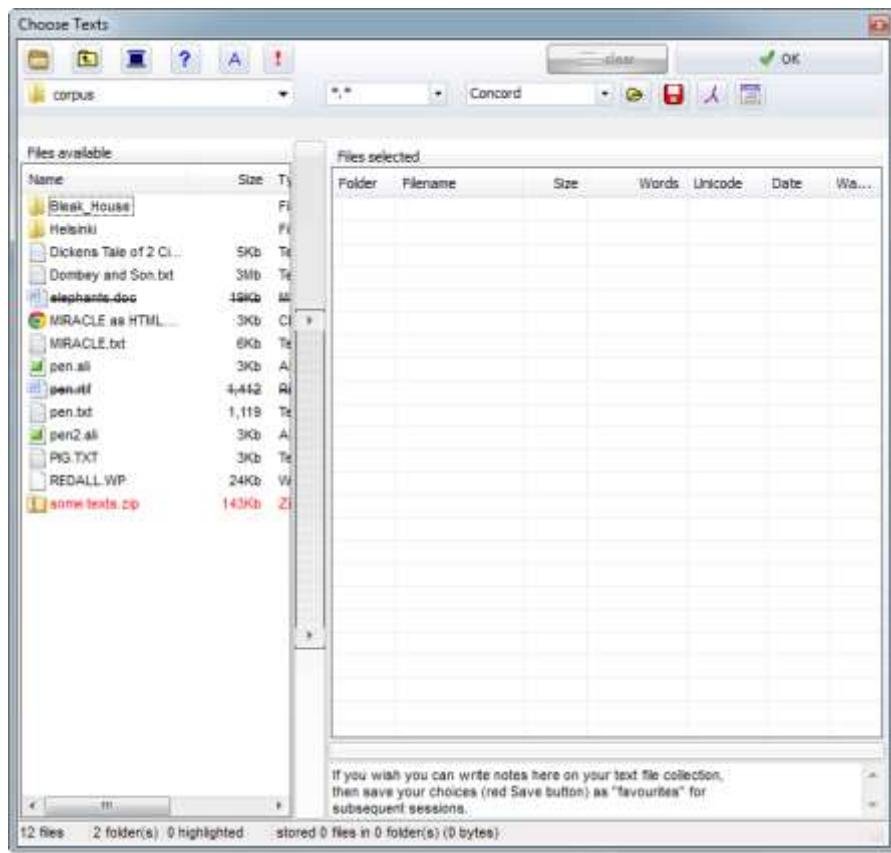
<그림 2> 텍스트 선택하기

그리면 <그림 3>과 같이 어떤 텍스트를 선택할 것인지를 묻는 팝업창이 나타난다.



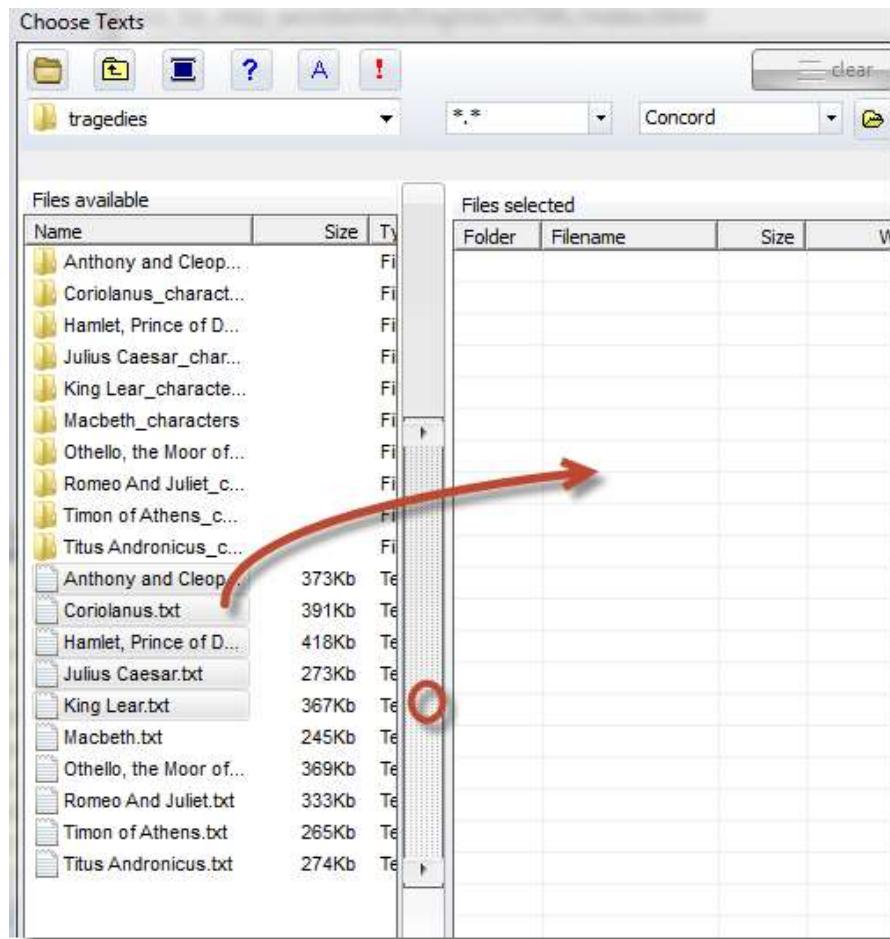
<그림 3> 텍스트 선택 팝업창

이 상태에서 *Choose Texts* 버튼을 클릭하면 <그림 4>와 같은 새로운 팝업창이 나타난다.



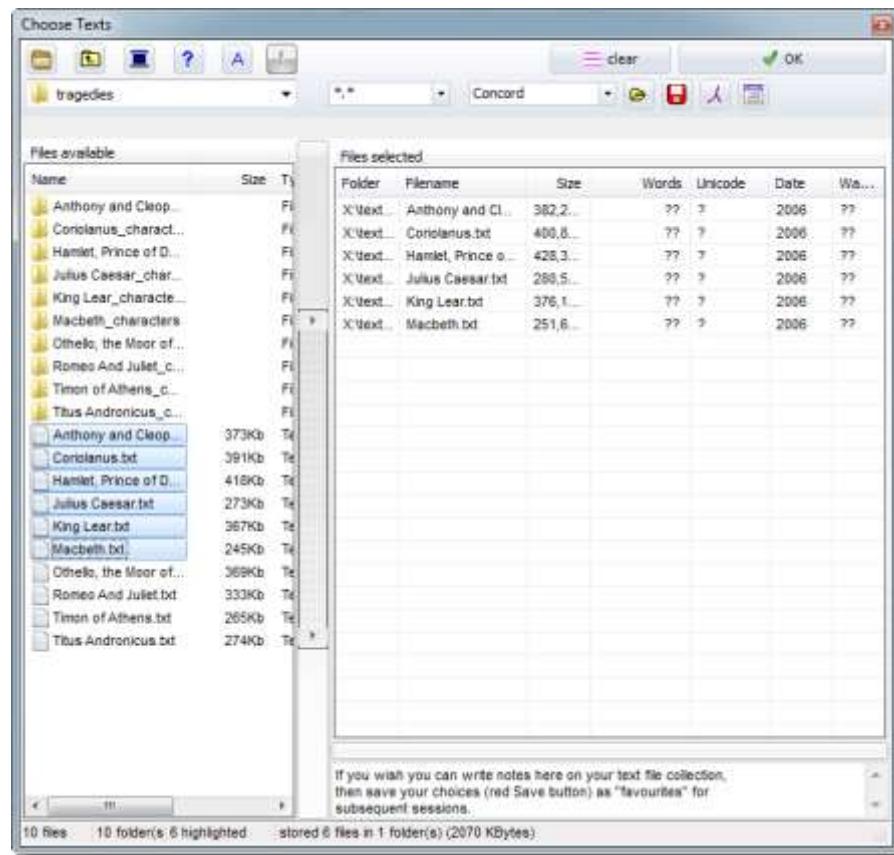
<그림 4> 텍스트 선택하기

<그림 4>에서 왼쪽은 우리가 흔히 보는 파일 탐색기이고, 오른쪽은 사용자가 선택한 텍스트 파일이 최종적으로 저장되는 곳이다. 사용자가 분석 대상으로 삼은 텍스트가 현재 다른 폴더에 있다면, 폴더 아이콘 ( )을 클릭하여 해당 텍스트가 있는 곳으로 이동한다. 참고로 워드스미스에서 사용하는 텍스트는 확장자의 파일 형식이 (.txt)로 되어있는 플레인 텍스트이어야 한다.



<그림 5> 선택한 텍스트 이동하기

사용자가 분석하려는 텍스트를 찾아 <그림 5>와 같이 해당 텍스트 파일을 선택한 다음, 그림 중간에 있는 막대를 클릭하거나 선택한 텍스트 파일을 오른쪽으로 드래그한다. 그러면 <그림 6>과 같이 사용자가 선택한 파일이 오른쪽으로 이동한다.



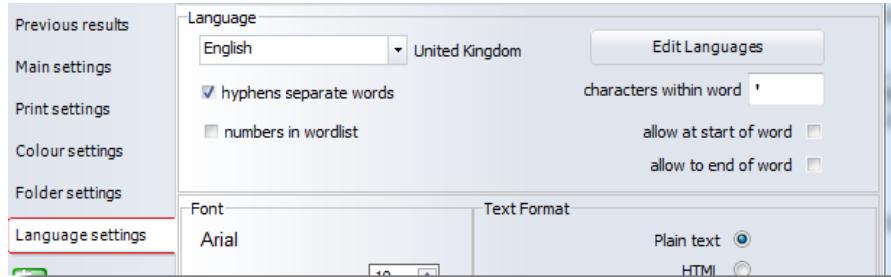
<그림 6> 선택한 텍스트 이동 완료

<그림 6>은 사용자가 콘코드(Concord)를 하기 위해 선택한 6 개의 파일이 오른쪽에 저장된 모습이다. (참고로 콘코드 바로 아래에 선택한 파일(Files selected)이 있음).

이제 버튼을 누르거나, 오른쪽 위에 있는 빨간색 x 박스를 클릭하여 창을 닫는다.

### 3. 다른 언어 선택하기

이 튜토리얼에서는 영어로 된 텍스트를 샘플로 다룬다. 한국어나 다른 언어로 된 텍스트를 분석하려면, <그림 7> 같이 콘트롤러에서 사용자가 원하는 언어를 다시 선택해야 한다.



<그림 7> 언어 선택

언어를 다시 선택하려면, <그림 8>과 같이 언어(Language)에서 드롭 다운하여 원하는 언어를 선택한다.



<그림 8> 언어 선택 메뉴

만일 드롭 다운 리스트에 원하는 언어가 없으면, *Edit Languages*를 클릭한 다음, 그곳에서 사용자가 원하는 언어를 찾는다(<그림 9>).



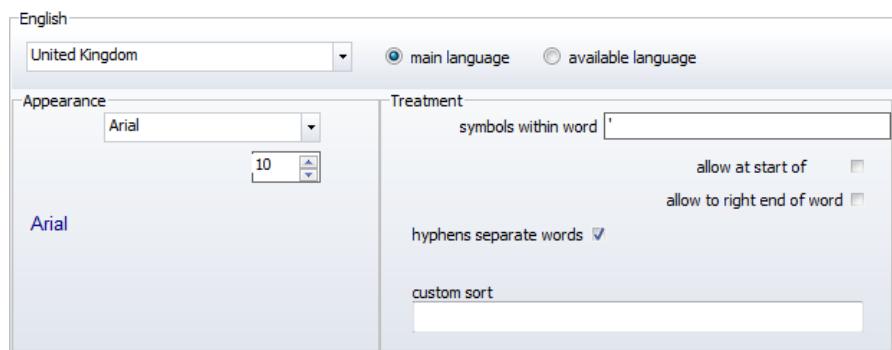
<그림 9> *Edit Languages*에서 언어 선택

사용자가 원하는 언어를 찾으면, 해당 언어를 클릭하여 선택한 다음 오른쪽으로 드래그하거나 중간에 있는 막대를 클릭한다(<그림 10>).

Name	Variety	S...	Font	S...	Word-letters
Arabic	Saudi Arabia	*	Arial	10	-
Chinese (Si...	People's R...	*	Arial	10	-
English	United King...	#1	Arial	10	'
French	France	*	Arial	10	-
German	Germany	*	Arial	10	-
Konkani	India		Arial	10	-
Igbo	Nigeria	*	Taho...	8	-

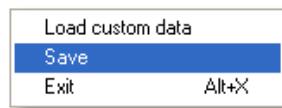
<그림 10> 사용자가 선택한 언어

참고로 이 설정에서 사용자가 원하는 언어를 기본 언어(워드스미스가 시작할 때 사용하는 기본 언어)로 설정하거나, 사용 가능한 언어로 설정을 변경할 수 있는 옵션이 있다.



<그림 11> 기본 언어 설정

<그림 11>은 기본 언어가 영어이고, 아포스트로피가 들어간 단어는 1 개로 처리하고, *self-conscious* 와 같이 하이픈이 들어간 단어는 2 개로 처리하고, 기본 글꼴은 *Arial 10*으로 하는 것으로 설정한 상태이다. 그럼 이제 지금까지 설정한 내용을 저장한다(<그림 12>).



<그림 12> 설정 저장

## 4. 콘코던싱

### 4.1 개요

<그림 13>은 워드스미스로 콘코던스한 결과이다.

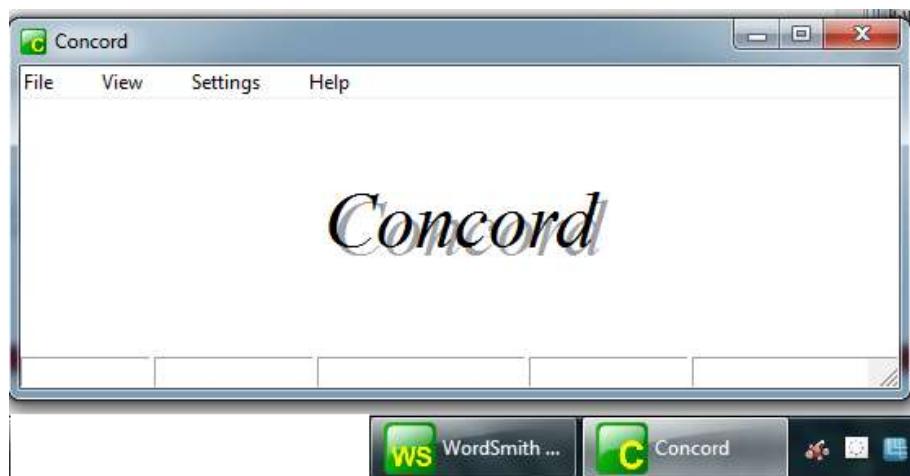


<그림 13> 콘코던스 결과

<그림 13>은 로미오와 줄리엣에서 **wherfore** 가 사용된 콘코던스 라인을 보여준다. 위 결과에 따르면, 로미오와 줄리엣에서 **wherfore** 가 총 5 회 사용된 것으로 나타났다.

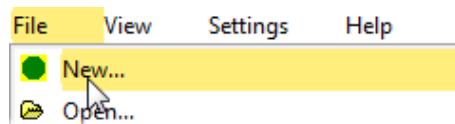
## 4.2 콘코드 만들기

워드스미스 콘트롤러에서 콘코드(Concord) 버튼을 클릭하면, <그림 14>와 같은 콘코드 툴(Concord Tool, 이하 '콘코드') 팝업창이 나타난다.



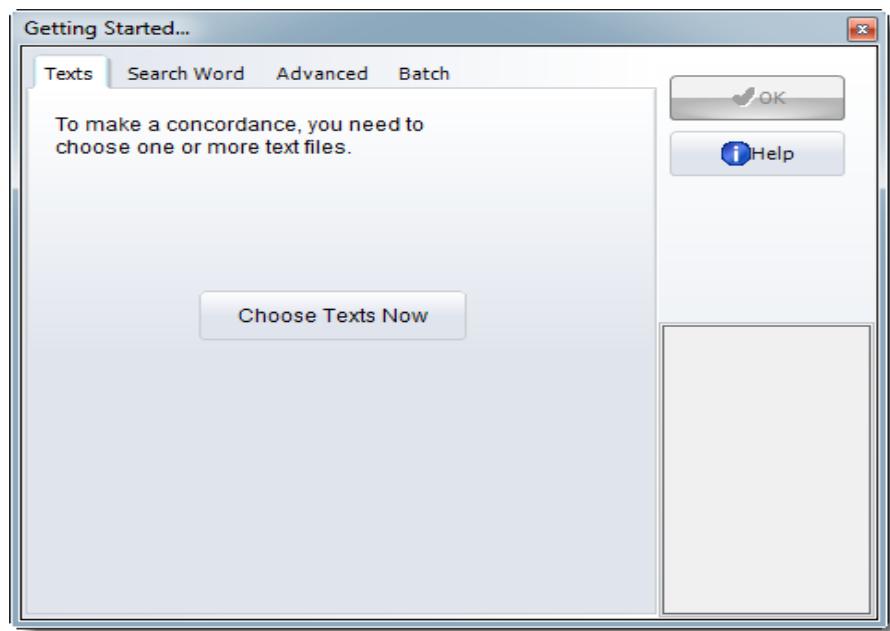
<그림 14> 콘코드 팝업창

메뉴에서 *File / New*를 선택한다(<그림 15>).



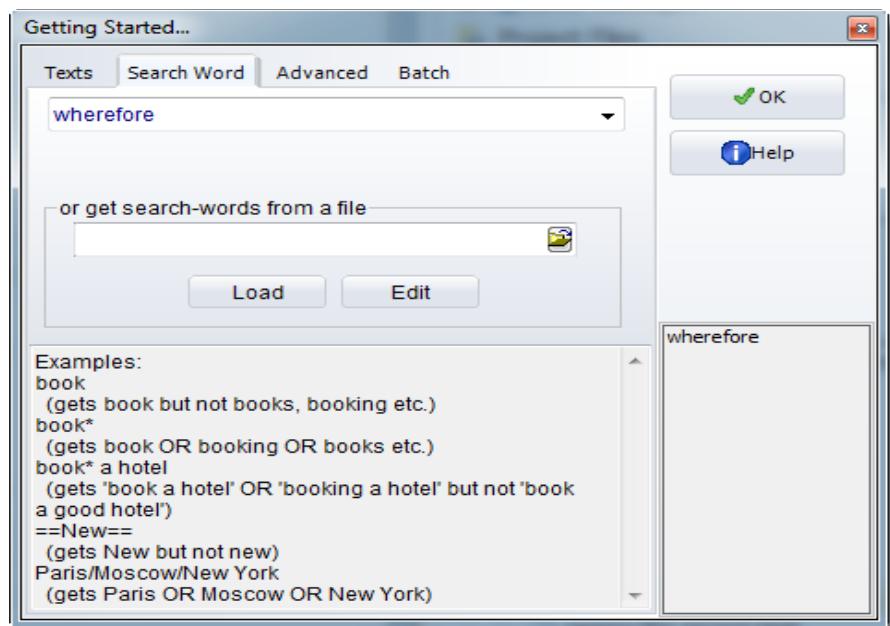
<그림 15> 콘코드 시작하기

만일 사용자가 아직 콘코드할 텍스트 파일을 선택하지 않았다면, <그림 16> 팝업창에서 중간에 있는 *Choose Texts Now* 버튼을 클릭하여 사용자가 원하는 텍스트를 선택한다.



<그림 16> 사용자가 원하는 텍스트 선택하기

사용자가 원하는 텍스트를 모두 선택한 다음 <그림 17> 팝업 창에서 *Search Word* 탭 아래 있는 빈 칸에 원하는 단어를 입력한다.



<그림 17> 사용자가 원하는 단어를 입력하기

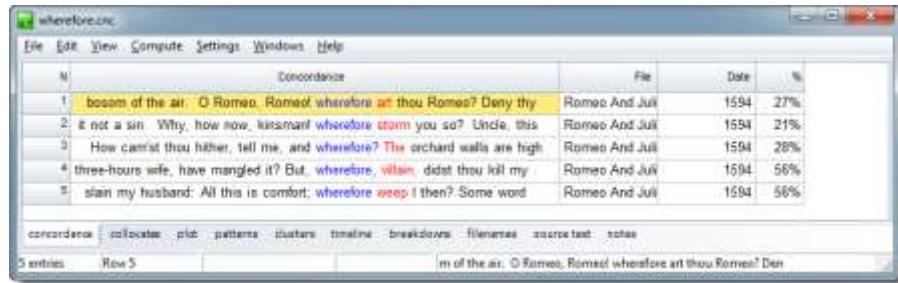
참고로 이 튜토리얼에서는 검색할 단어로 **wherefore** 를 선택하였다. 이제 **OK** 버튼을 누른다. <그림 18>에서 볼 수 있듯이, 콘코던스는 구두점이나 공백과 같은 단어 분리 기호로 **wherefore** 를 다른 단어와 분리한 다음 로미오와 줄리엣에서 사용된 모든 **wherefore** 의 사용 예를 보여준다.



<그림 18> 로미오와 줄리엣에서 **wherefore** 사용 예

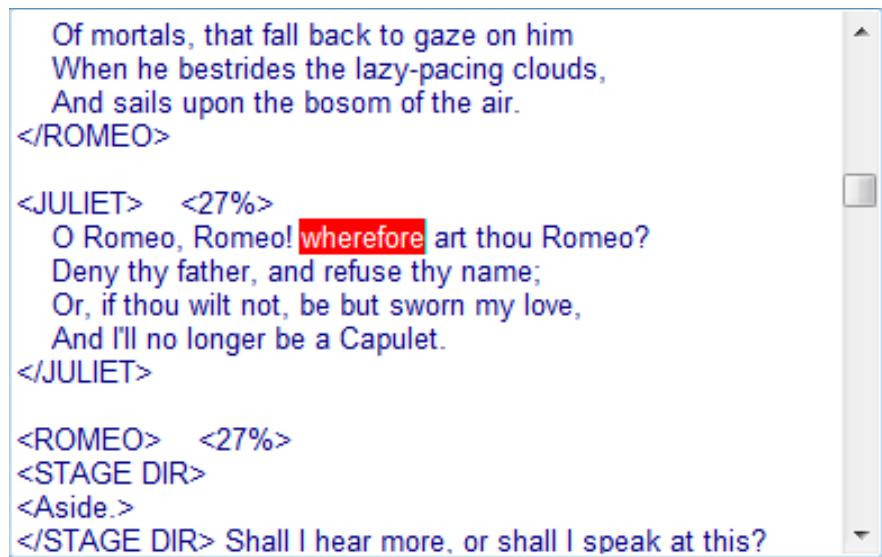
#### 4.3 소스 텍스트 보기

콘코던스 결과에서 소스 텍스트를 보려면, <그림 19>에 있는 콘코던스 라인에서 사용자가 원하는 콘코던스 라인을 더블 클릭한다.



<그림 19> 콘코던스 라인

<그림 20>은 소스 텍스트를 보기 위해 *wherfore art thou Romeo* 를 포함한 콘코던스 라인을 더블 클릭한 결과 모습이다.



<그림 20> 콘코던스 라인을 더블 클릭하여 소스 텍스트 보기

이상과 같은 방식으로 소스 텍스트를 보는 방법 이외에, 사용자가 원하는 콘코던스 라인을 선택한 상태에서 F8 을 누르면 <그림 21>과 같이 해당 콘코던스 라인이 영역이 넓어진다.

N	Concordance	Set	Tag	Wo
1	to gaze on him When he bestrides the lazy-pacing clouds, And sails upon the bosom of the air. O Romeo, Romeo! <b>wherefore art thou Romeo?</b> Deny thy father, and refuse thy name; Or, if thou wilt not, be but sworn my love, And I'll no			5.
	expand this line by pulling it down			
	it not a sin. Why, how now, kinsman! <b>wherefore storm</b> you so? Uncle, this is a Montague, our foe; A villain that is hither come in spite, To scorn at our			4.
3	Art thou not Romeo, and a Montague? Neither, fair maid, if either thee dislike. How cam'st thou hither, tell me, and <b>wherefore?</b> The orchard walls are high and hard to climb, And the place death, considering who thou art, If any of my			5.
4	my husband? Ah! poor my lord, what tongue shall smooth thy name, When I, thy three-hours wife, have mangled it? But, <b>wherefore, villain,</b> didst thou kill my cousin? That villain cousin would have kill'd my husband: Back, foolish tears,			11.
5	husband lives, that Tybalt would have slain: And Tybalt's dead, that would have			11.

<그림 21> F8 을 눌러 소스 텍스트 보기

만일 해당 콘코던스 라인을 더 자세히 들여다보고 싶으면, 마우스 커서를 왼쪽 가장자리에 있는 번호 칼럼으로 이동하여 마우스 커서를 번호 칼럼의 아래쪽 라인을 클릭한 상태에서 아래로 내리면, 콘코던스 라인이 <그림 22>와 같이 넓어진다.

N	Concordance	Set	Tag	Wo
1	to gaze on him When he bestrides the lazy-pacing clouds, And sails upon the bosom of the air. O Romeo, Romeo! <b>wherefore art thou Romeo?</b> Deny thy father, and refuse thy name; Or, if thou wilt not, be but sworn my love, And I'll no			5.
2	expand this line by pulling it down			
	it not a sin. Why, how now, kinsman! <b>wherefore storm</b> you so? Uncle, this is a Montague, our foe; A villain that is hither come in spite, To scorn at our			4.

<그림 22> 콘코던스 라인 영역 확장하기

만일 콘코던스 라인의 폭을 더 넓히고 싶으면, 위쪽에 있는 Set 칼럼의 왼쪽 라인으로 마우스 커서를 이동하여 클릭한 상태에서 오른쪽으로 이동하면, <그림 23>과 같이 콘코던스 라인의 폭을 더 넓힐 수 있다

N	Concordance	S
1	I might touch that cheek. Ay me! She speaks: O! speak again, bright angel; for thou art As glorious to this night, being o'er my head, As is a winged messenger of heaven Unto the white-upturned wond'ring eyes Of mortals, that fall back to gaze on him When he bestrides the lazy-pacing clouds, And sails upon the bosom of the air. O Romeo, Romeo! <b>wherefore art thou Romeo?</b> Deny thy father, and refuse thy name; Or, if thou wilt not, be but sworn my love, And I'll no longer be a Capulet. Shall I hear more, or shall I speak at this? Tis but thy name that is my enemy; Thou art thyself though, not a Montague. What's Montague? it is nor hand, nor foot, Nor arm, nor face, nor any other part	
2	solemnity? Now, by the stock and honour of my kin, To strike him dead I hold	

<그림 23> 콘코던스 라인 영역 추가 확장해서 보기

#### 4.4 콜로케이트와 상호정보

<그림 24>는 BNC의 문어 파트에서 Ago의 콜로케이트 결과를 빈도순서대로 정렬한 화면이다.

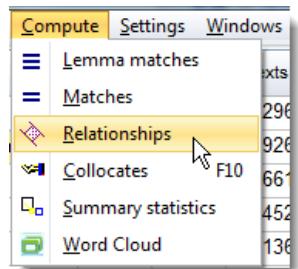
N	Word	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	
1	AGO	2,296	16,779	31	31	9	12	9	0	1	16,717	
2	YEARS	1,926	8,994	8,938	56	14	12	20	3	6,889	0	
3	A	1,661	6,312	4,525	1,787	357	669	1,589	1,910	0	0	2
4	THE	1,452	4,518	1,361	3,157	747	533	47	25	9	0	7
5	WAS	1,136	2,465	1,150	1,315	323	490	294	43	0	0	1
6	OF	1,092	2,372	1,205	1,167	310	220	234	441	0	0	0
7	AND	1,103	2,315	536	1,779	165	203	139	28	1	0	1,1
8	TWO	849	2,102	2,019	83	48	56	77	1,806	32	0	
9	TO	986	1,999	561	1,438	262	229	66	3	1	0	3
10	IN	885	1,755	711	1,044	311	344	51	5	0	0	3
11	LONG	820	1,558	1,524	34	8	9	8	345	1,154	0	
12	THAT	802	1,481	698	783	319	208	112	58	1	0	2
13	MONTHS	616	1,372	1,362	10	1	2	3	0	1,356	0	
14	I	589	1,259	405	854	145	236	22	2	0	0	5
15	IT	728	1,183	604	579	230	237	98	39	0	0	2
16	THREE	602	1,158	1,112	46	34	36	31	1,011	0	0	
17	HAD	610	1,129	431	698	179	128	65	59	0	0	
18	SOME	593	1,068	985	83	31	54	179	721	0	0	
19	FEW	639	1,067	1,039	28	3	4	16	1,016	0	0	
20	WEEKS	470	1,038	1,029	9	4	1	1	0	1,023	0	
21	YEAR	543	1,016	964	52	9	20	31	1	903	0	
22	WE	807	994	717	801	800	97	17	0	0	0	

concordance    collocates    plot    patterns    clusters    timeline    breakdowns    filenames    source text    notes

1,125 entries    Row 2    YEARS

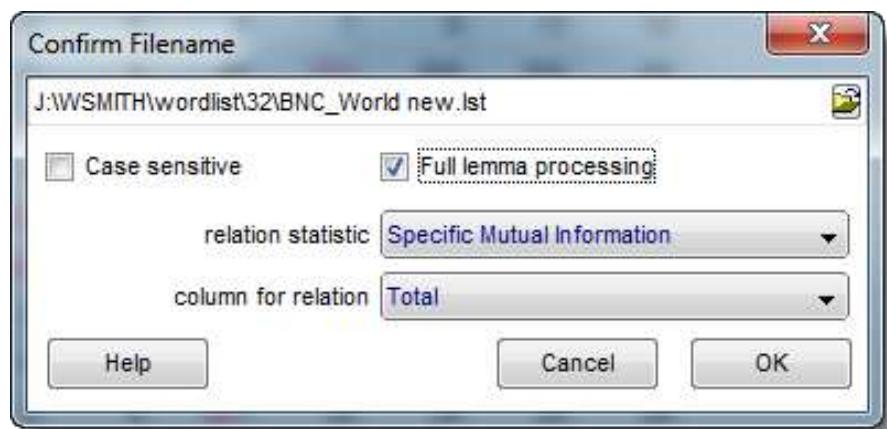
<그림 24> Ago 콜로케이트 결과

<그림 24>에서 볼 수 있듯이, BNC 문어 파트에서 Ago는 약 17,000 번 사용되었다. Years는 Ago와 가장 강한 콜로케이트 관계를 보여주며, Ago 인근에서 함께 약 9,000 번 사용된 것으로 나타났다. 이러한 결과를 바탕으로 Years는 Ago와 중요한 콜로케이트 관계가 있음을 알 수 있다. 참고로 <그림 24>를 자세히 들여다보면, 대부분이 Ago의 L1(왼쪽 첫 번째 위치)에서 주로 사용된 것을 확인할 수 있다. 또한, <그림 24>를 분석하면 The가 Long, Month, Weeks, 등 보다 빈도수가 더 높은 것을 알 수 있다. 이밖에 다른 단어도 Ago와 콜로케이트 관계가 있는 것으로 나타났는데, 그렇다면 A, The, Was와 같은 단어들도 Ago와 정말 가까운 콜로케이트 관계가 있는 것인지에 관한 물음에 고민을 해 볼 필요가 있다. 이 질문에 대한 대답은 다음과 같은 과정을 통해 확인할 수 있다. 메뉴에서 *Compute / Relationships*를 선택한다(<그림 25>).



<그림 25> Compute / Relationships 기능 선택하기

그리면 <그림 26>과 팝업창이 나타나는데, Ago 와의 콜로케이트 결과를 비교하기 적절한 워드리스트를 불러온 다음 OK 버튼을 클릭한다.



<그림 26> 상호정보

이상과 같은 과정을 거치면 Ago 의 콜로케이트 결과가 <그림 27>과 같이 새롭게 나타난다. <그림 27>은 Ago 의 새로운 콜로케이트 결과를 화면 위쪽에 있는 *Relation* 칼럼을 선택하여 새롭게 콜로케이트 결과를 정렬한 것이다.

N	Word	Relation ↓	Texts	Total	To L
1	AGO	12.126	2,296	16,779	
2	YEARS	9.028	1,926	8,994	8,9
3	FORTNIGHT	8.894	77	121	1
4	MOONS	8.644	13	13	
5	WEEKS	8.470	470	1,038	1,0
6	SEASONS	8.440	41	81	
7	MOMENTS	8.248	20	178	1
8	MONTHS	8.179	616	1,372	1,3
9	UNTHINKABLE	8.042	18	18	
10	DECADE	7.843	147	164	1
11	CENTURIES	7.491	102	124	1
12	COUPLE	7.311	252	358	3
13	DECades	7.008	54	61	

◀
▶

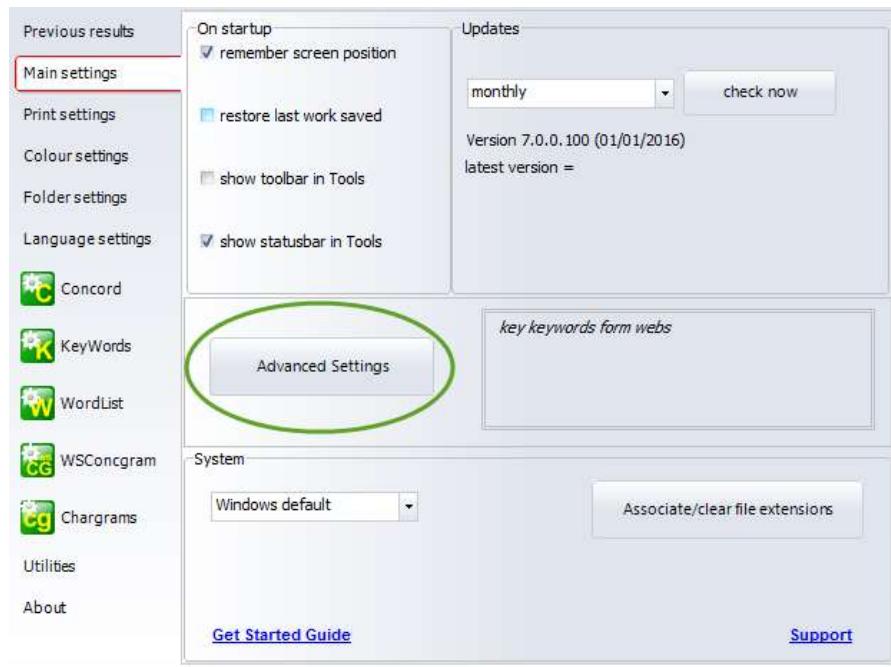
[concordance](#)
collocates
[plot](#)
[patterns](#)
[clusters](#)
[timeliner](#)

<그림 27> 새로운 Ago 콜로케이트 결과

<그림 27>을 자세히 살펴보면, Ago 와 콜로케이트 관계를 보이는 리스트의 상위권 단어들이 주로 기간이나 숫자이며, 이러한 결과를 통해 어떤 단어들이 Ago 와 함께 자주 사용되는 경향이 있는지를 파악할 수 있다.

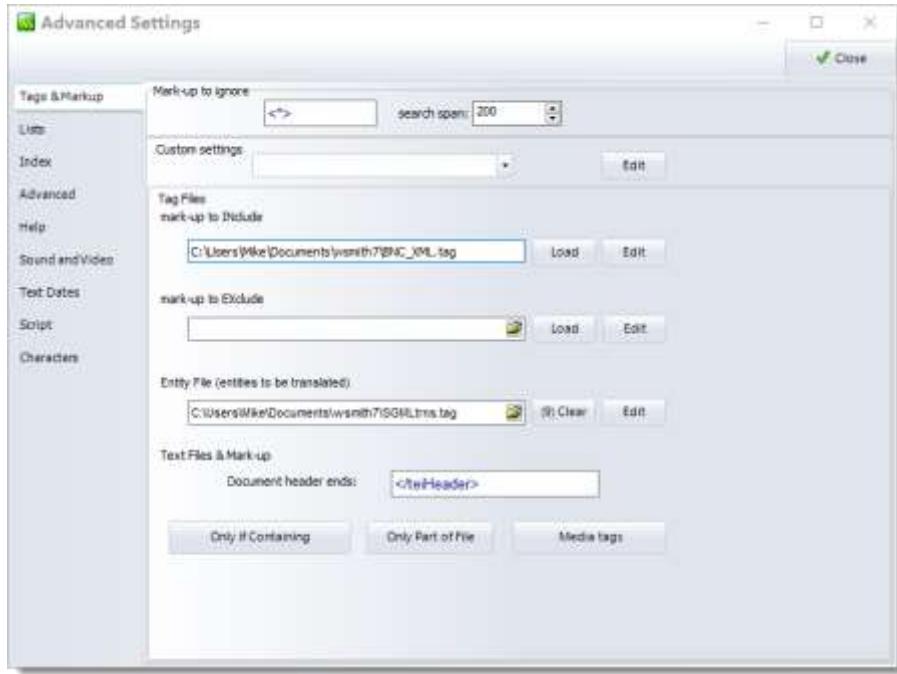
#### 4.5 태그된 텍스트 콘코던싱(1)

사용자가 분석하고자 하는 소스 텍스트가 태그된 상태라면, 사용자는 이런 사실을 워드스미스에 미리 알려주어야 한다. 이를 위해서는 콘트롤러에서 *Advanced Settings* 버튼을 클릭한다(<그림 28>).



<그림 28> *Advanced Settings* 선택하기

그러면 <그림 29>와 같은 팝업창이 나타나는데, 이곳에서 *Tags & Markup* 탭을 선택한다.



<그림 29> *Tags & Markup* 탭 선택하기

<그림 29>는 워드스미스에 꺽쇠 괄호(<>)로 시작하고 끝나는 모든 태그를 무시하고, 몇 가지 개체 참조를 % 나 "와 같은 기호로 변환하고, 각 텍스트의 헤더를 (</teiHeader>)태그가 나오면 자르도록 설정한 상태이다. 그럼 이제 워드스미스를 사용할 준비가 어느 정도 끝났다.

#### 4.6 태그된 텍스트 콘코던싱(2)

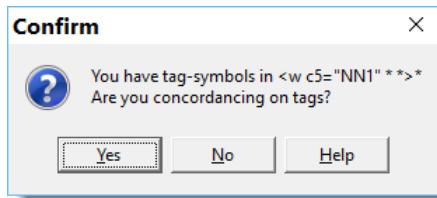
그럼 이제는 품사를 콘코던스 해본다. BNC에서는 아래와 같이 단수 가산 명사를 마크업하였다.

<w c5="NN1"...>

만일 BNC에서 사용된 단수 가산 명사를 모두 찾고 싶다면, 아래와 같이 입력한다.

<w c5="NN1" \* \*>\*

그러면 <그림 30>과 같이 꺽쇠 괄호가 텍스트에 사용된 문자인지, 아니면 태그가 시작되거나 끝나는 표시인지를 묻는 팝업창이 나타난다.



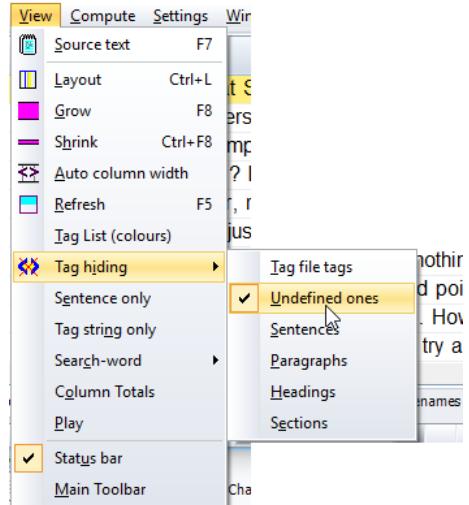
<그림 30> 태그된 품사를 검색하는지 묻는 팝업창

여기서 Yes를 선택하면, <그림 31>과 같은 결과가 나타난다.



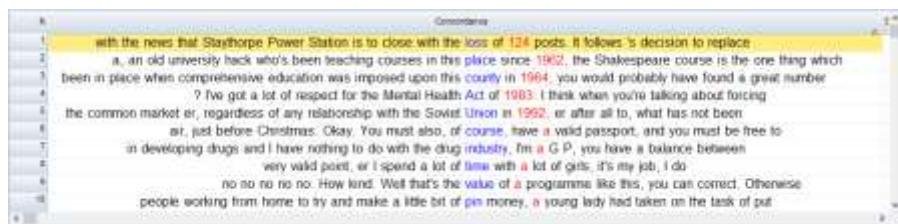
<그림 31> 태그된 텍스트 콘코던스 라인

만일 사용자가 정의되지 않은 태그의 결과 화면을 보고 싶지 않으면, 메뉴에서 View를 선택한 다음 *Undefined ones*를 클릭한다 (<그림 32>).



<그림 32> 태그 표시를 화면에서 보지 않기

그러면 <그림 33>과 같이 태그 표시가 사라진 화면으로 바뀐다.

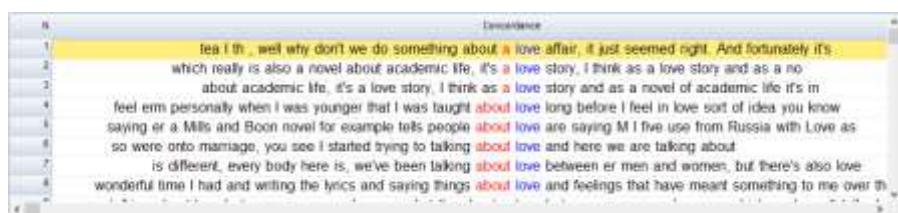


<그림 33> 태그 표시가 없는 결과

만일 사용자가 아래와 같이 입력하면,

<w c5="NN1" \* \*>love

BNC에서 love가 동사로 사용된 경우가 아니라, <그림 34>와 같이 단수 가산 명사로 사용된 경우만 보여준다.



<그림 34> love의 단수 가산 명사만 보기

## 5. 워드리스트

### 5.1 개요

<그림 35>는 워드스미스로 워드 리스트를 추출한 것이다.

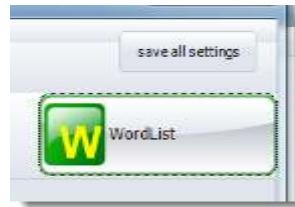
N	Word	Freq.	%	Texts	%
6,231	WHEN	107	0.20	2	100.00
6,232	WHENCE	3		2	100.00
6,233	WHENSOEVER	1		1	50.00
6,234	WHER	3		1	50.00
6,235	WH'ER	1		1	50.00
6,236	WHERE	106	0.20	2	100.00
6,237	WHEREAT	1		1	50.00
6,238	WHEREFORE	8	0.01	2	100.00
6,239	WHEREIN	9	0.02	2	100.00
6,240	WHEREOF	2		1	50.00
6,241	WHEREON	6	0.01	2	100.00
6,242	WHERE'S	8	0.01	2	100.00
6,243	WHERETO	5		2	100.00

<그림 35> 워드 리스트

워드 리스트는 텍스트에서 각 단어가 몇 번 반복되어 사용되었는지, 텍스트에서 각 단어가 사용된 비율, 각 단어가 사용된 텍스트 파일 수 등을 보여준다.

## 5.2 워드 리스트 만들기

워드 리스트를 만들려면, 우선 [콘트롤러](#)에서 워드리스트(WordList) 버튼을 클릭한다(<그림 36>).



<그림 36> 워드리스트 기능 선택하기

워드리스트(WordList) 버튼을 클릭하면, <그림 37>과 같이 사용자가 만들고자 하는 워드리스트의 소스 텍스트를 선택하라는 팝업창이 나타난다.



<그림 37> 소스 텍스트 선택하기

이 튜토리얼에서는 2개의 텍스트 파일(로미오와 줄리엣과 햄릿)으로 간단한 워드 리스트를 만들어 본다. 2개의 텍스트 파일을 선택하여 모든 준비가 끝나면, *Make a word list now* 버튼을 클릭한다.

워드 리스트는 기본적으로 단어의 빈도를 보여준다. <그림 38>에서 볼 수 있듯이, 로미오와 줄리엣과 햄릿을 합친 2개의 텍스트에서 가장 자주 사용된 단어는 The, And, To 등이다. 각 단어 옆에는 2개의 텍스트에서 해당 단어가 얼마나 자주 사용되었는지에 대한 빈도수와 각각의 단어가 2개의 텍스트에서 사용된 비율 및 해당 단어가 2개의 텍스트에 나타난 비율을 확인할 수 있다.

The screenshot shows a Windows application window titled "hamlet and romeo.lst". The menu bar includes File, Edit, View, Compute, Settings, Windows, and Help. The main window displays a table of word frequencies:

N	Word	Freq.	%	Texts	% Lemmas	Set
1	THE	1,757	3.26	2	100.00	
2	AND	1,565	2.90	2	100.00	
3	TO	1,298	2.41	2	100.00	
4	I	1,141	2.12	2	100.00	
5	OF	1,044	1.94	2	100.00	
6	A	970	1.80	2	100.00	
7	MY	875	1.62	2	100.00	
8	YOU	846	1.57	2	100.00	
9	IN	742	1.38	2	100.00	
10	THAT	741	1.37	2	100.00	
11	IS	690	1.28	2	100.00	
12	IT	640	1.19	2	100.00	
13	NOT	574	1.06	2	100.00	
14	THIS	512	0.95	2	100.00	

Below the table, there is a navigation bar with tabs: frequency (selected), alphabetical, statistics, filenames, and notes. At the bottom, status bars show "6,463 entries", "Row 1", and "YOU".

<그림 38> 로미오와 줄리엣과 햄릿 워드 리스트

워드 리스트는 기본적으로 빈도수가 높은 순서대로 정렬된다. 만일 알파벳 순서대로 워드 리스트를 보고 싶다면, <그림 39>와 같이 왼쪽 아래에 있는 *alphabetical* 탭을 선택한다.

N	Word	Freq.	%	Texts	%
6,231	WHEN	107	0.20	2	100.00
6,232	WHENCE	3		2	100.00
6,233	WHENSOEVER	1		1	50.00
6,234	WHE'R	3		1	50.00
6,235	WH'E'R	1		1	50.00
6,236	WHERE	106	0.20	2	100.00
6,237	WHEREAT	1		1	50.00
6,238	WHEREFORE	8	0.01	2	100.00
6,239	WHEREIN	9	0.02	2	100.00
6,240	WHEREOF	2		1	50.00
6,241	WHEREON	6	0.01	2	100.00
6,242	WHERE'S	8	0.01	2	100.00
6,243	WHERETO	5		2	100.00

frequency alphabetical statistics filenames notes

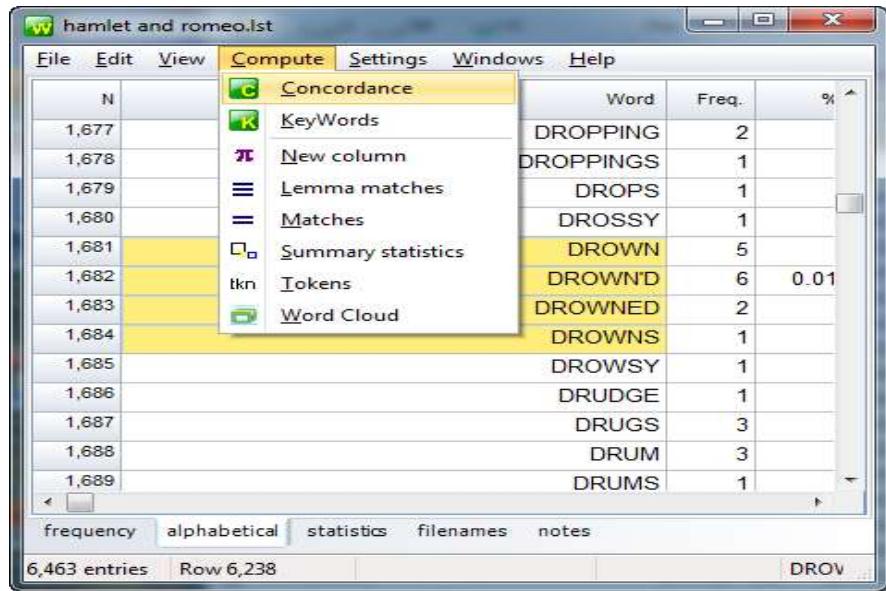
6,463 entries Row 1 WHE

<그림 39> 알파벳 순서대로 워드 리스트 정렬하기

그럼 이제 워드 리스트에서 *wherefore* 가 있는 화면 아래쪽으로 스크롤 해보자. <그림 39>에서 볼 수 있듯이, *wherefore* 는 로마오와 줄리엣과 햄릿을 합친 2 개의 텍스트에서 총 8 번 사용되었고, 전체 텍스트에서 사용된 비율은 0.01%이며, 2 개의 텍스트에서 모두 사용된 것으로 나타났다.

### 5.3 선택한 단어 콘코던싱

워드 리스트를 추출하고 난 후, 특정 단어가 어떤 문맥에서 사용되었는지를 살펴볼 수 있다. 워드 리스트에서 특정 단어를 콘코던싱 하고 싶으면 다음과 같이 한다. 우선 관심있는 단어를 1개, 또는 그 이상 선택한다(<그림 40>).



<그림 40> 관심있는 단어 선택하기

메뉴에서 *Compute / Concordance*를 선택한다(<그림 40>). 그러면 <그림 41>과 같은 결과 화면이 나타난다.

N	Concordance	Set
1	tears to fires! And these, who often drown'd could never die, Transparent	DROW'ND
2	lay To muddy death. Alas! then, she is drown'd? Drown'd, drown'd. Too much	DROW'ND
3	death. Alas! then, she is drown'd? Drown'd, drown'd. Too much of water	DROW'ND
4	act, to do, and to perform: argal, she drowned herself wittingly. Nay, but	DROWNED
5	burial. How can that be, unless she drowned herself in her own defence?	DROWNED
6	that? but if the water come to him, and drown him, he drowns not himself.	DROWN
7	good: if the man go to this water, and drown himself, it is, will he, nill he, he	DROWN
8	heel, So fast they follow: your sister's drown'd, Laertes. Drown'd! O, where?	DROW'ND
9	be else. For here lies the point: if I drown myself wittingly it argues an act;	DROWN
10	water come to him, and drown him, he drowns not himself. argal, he that is	DROWNS
11	follow: your sister's drown'd, Laertes. Drown'd! O, where? There is a willow	DROW'ND
12	have countenance in this world to drown or hang themselves more than	DROWN
13	cue for passion That I have? He would drown the stage with tears, And cleave	DROWN
14	Alas! then, she is drown'd? Drown'd, drown'd. Too much of water hast thou,	DROW'ND

<그림 41> 여러 단어 콘코던싱 결과

## 5.4 레마타이징

워드 리스트 결과 화면에서 수동으로 레마타이징하려면, 레마타이징하려는 단어를 선택한다(<그림 42>)

BLEEDING		
571	BLEEDS	1
572	BLEENCH	1
573	BLESS	5
574	BLESS'D	4
575	BLESSED	5
576	BLESSING	6
577	BLESSINGS	1
578	BLIND	6
579	BLISS	2

<그림 42> 레마타이징 대상 단어 선택하기

그런 다음, 왼쪽 마우스 버튼을 눌러 드래그하여 사용자가 합치고 싶은 라인으로 마우스 커서를 이동한다(<그림 43>).

N	Word	Freq.	%	Texts
569	BLEED	2		1
570	BLEEDING	3		2 1
571	BLEEDS	1		1
572	BLEENCH	1		1
573	BLESS	5		2 1
574	BLESS'D	4		2 1
575	BLESSED	5		1
576	BLESSING	6	0.01	2 1
577	BLESSINGS	1		1
578	BLIND	6	0.01	2 1
579	BLISS	2		1
580	BLISTER	1		1

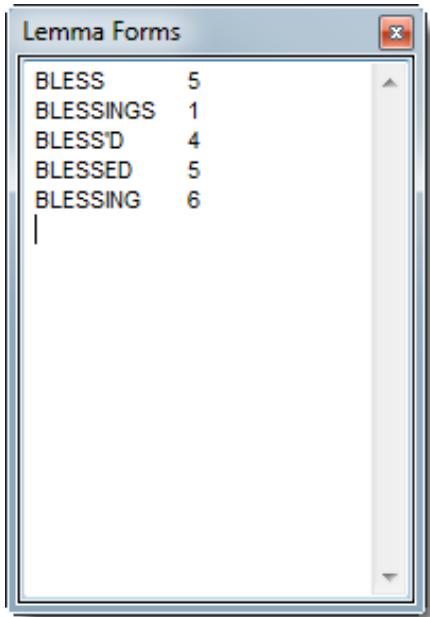
<그림 43> 마우스 커서 이동하기

그런 다음 왼쪽 마우스 버튼을 놓는다<그림 44>).

572	BLEND	1	1	50.00	
573	BLESS	21	2	100.00	bless[5] blessings[1] ble
574	BLESS'D	4	2	100.00	
575	BLESSED	5	1	50.00	
576	BLESSING	6	0.01	2	100.00
577	BLESSINGS	1	1	50.00	
578	BLIND	6	0.01	2	100.00
579					

<그림 44> 레마타이징

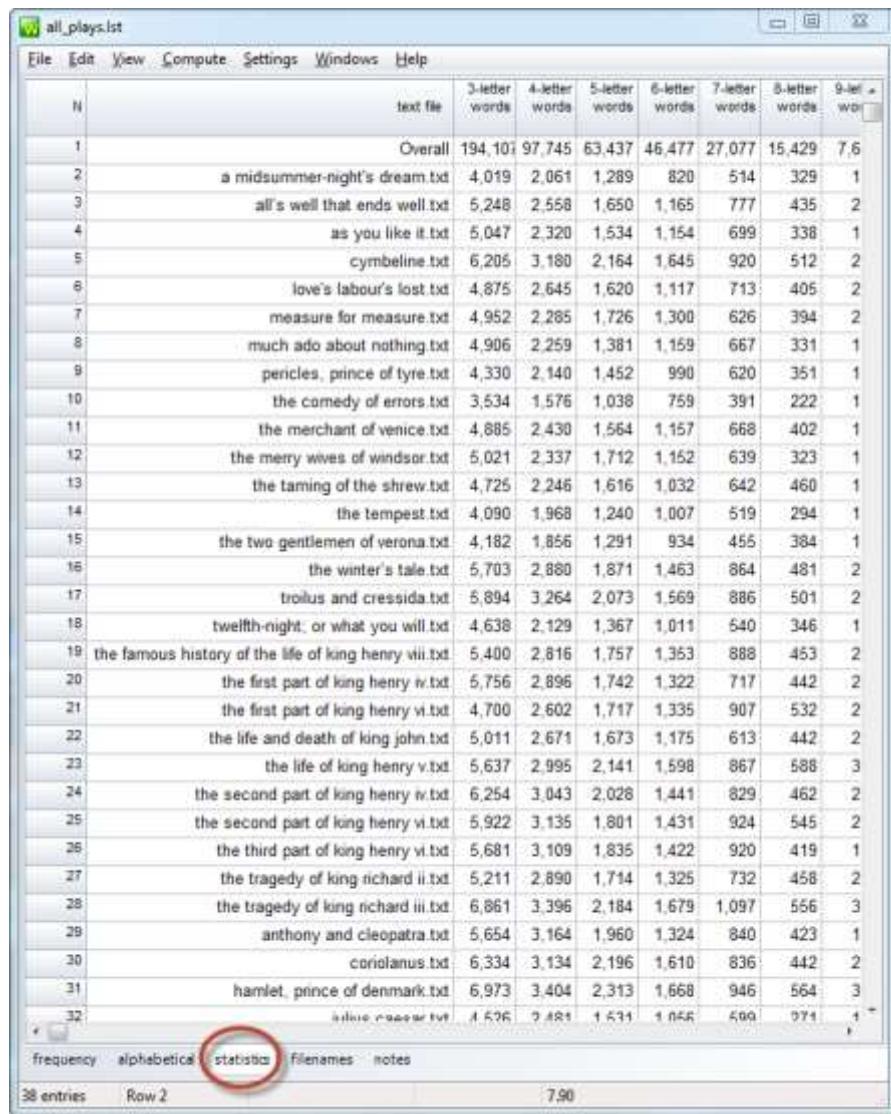
그러면 <그림 44>와 같이 빈도수가 바뀌고, 레마(Lemmas) 칼럼에 해당 단어가 보이게 된다. 단어가 너무 많은 경우는, 레마 칼럼을 더블 클릭하면 세부 정보를 볼 수 있다(<그림 45>).



<그림 45> 레마 품 보기

## 5.5 워드 리스트 통계

워드 리스트 아래에 있는 *statistics* 탭을 선택하면, <그림 46>과 같은 통계 결과가 나타난다.



The screenshot shows a software window titled "all\_plays.lst". The menu bar includes File, Edit, View, Compute, Settings, Windows, and Help. Below the menu is a table with columns: N, text file, 3-letter words, 4-letter words, 5-letter words, 6-letter words, 7-letter words, 8-letter words, 9-letter words, and 10-letter words. The table lists 38 entries, each corresponding to a Shakespeare play. The last row shows totals: 1,526, 2,821, 1,531, 1,046, 569, 571, 1. At the bottom of the table, there are tabs for frequency, alphabetical, statistics, filenames, and notes, with "statistics" being the active tab and circled in red. The status bar at the bottom indicates "38 entries Row 2 790".

N	text file	3-letter words	4-letter words	5-letter words	6-letter words	7-letter words	8-letter words	9-letter words	10-letter words
1	Overall	194,101	97,745	63,437	46,477	27,077	15,429	7,6	
2	a midsummer-night's dream.txt	4,019	2,061	1,289	820	514	329	1	
3	all's well that ends well.txt	5,248	2,558	1,650	1,165	777	435	2	
4	as you like it.txt	5,047	2,320	1,534	1,154	699	338	1	
5	cymbeline.txt	6,205	3,180	2,164	1,645	920	512	2	
6	love's labour's lost.txt	4,875	2,645	1,620	1,117	713	405	2	
7	measure for measure.txt	4,952	2,285	1,726	1,300	626	394	2	
8	much ado about nothing.txt	4,906	2,259	1,381	1,159	667	331	1	
9	pericles, prince of tyre.txt	4,330	2,140	1,452	990	620	351	1	
10	the comedy of errors.txt	3,534	1,576	1,038	759	391	222	1	
11	the merchant of venice.txt	4,885	2,430	1,564	1,157	668	402	1	
12	the merry wives of windsor.txt	5,021	2,337	1,712	1,152	639	323	1	
13	the taming of the shrew.txt	4,725	2,246	1,616	1,032	642	460	1	
14	the tempest.txt	4,090	1,968	1,240	1,007	519	294	1	
15	the two gentlemen of verona.txt	4,182	1,856	1,291	934	455	384	1	
16	the winter's tale.txt	5,703	2,880	1,871	1,463	864	481	2	
17	troilus and cressida.txt	5,894	3,264	2,073	1,569	886	501	2	
18	twelfth-night, or what you will.txt	4,638	2,129	1,367	1,011	540	346	1	
19	the famous history of the life of king henry viii.txt	5,400	2,816	1,757	1,353	888	453	2	
20	the first part of king henry iv.txt	5,756	2,896	1,742	1,322	717	442	2	
21	the first part of king henry vi.txt	4,700	2,602	1,717	1,335	907	532	2	
22	the life and death of king john.txt	5,011	2,671	1,673	1,175	613	442	2	
23	the life of king henry v.txt	5,637	2,995	2,141	1,598	867	588	3	
24	the second part of king henry iv.txt	6,254	3,043	2,028	1,441	829	462	2	
25	the second part of king henry vi.txt	5,922	3,135	1,801	1,431	924	545	2	
26	the third part of king henry vi.txt	5,681	3,109	1,835	1,422	920	419	1	
27	the tragedy of king richard ii.txt	5,211	2,890	1,714	1,325	732	458	2	
28	the tragedy of king richard iii.txt	6,861	3,396	2,184	1,679	1,097	556	3	
29	anthony and cleopatra.txt	5,654	3,164	1,960	1,324	840	423	1	
30	coriolanus.txt	6,334	3,134	2,196	1,610	836	442	2	
31	hamlet, prince of denmark.txt	6,973	3,404	2,313	1,668	946	564	3	
32	all the plays.txt	1,526	2,821	1,531	1,046	569	371	1	

frequency alphabetical statistics filenames notes

38 entries Row 2 790

<그림 46> 워드 리스트 통계 탭 선택하기

참고로 <그림 46>을 자세히 살펴보면, 셰익스피어 작품에서는 4 개 단어 뭉치보다는 3 개 단어 뭉치가 2 배 정도 더 많이 사용된 것을 발견할 수 있다.

## 5.6 멀티 워드 유닛

### 5.6.1 인덱스 사용하기

워드스미스를 사용하여 아래와 같은 2 개나 3 개 단어 뭉치(n-grams)를 만들려면, 사용자는 먼저 [index file](#) 을 먼저 만들어야 한다. 참고로 인덱스 파일은 텍스트 파일의 모든 단어 위치를 기록하여 각 단어의 어느 부분에 어떤 단어가 왔는지를 나중에 볼 수 있도록 하는 것이다.

OF THE

IN THE END

ONCE UPON A TIME

워드스미스에서는 인덱스 파일로 단어 유형을 서로 관련시키는 상호정보 수치를 계산하고, 단어 클러스터를 살펴볼 수 있다. 인덱스 파일은 콘그램(Congram) 검색을 생성하기 위해서도 필요하다.

참고: [making the multi-word unit wordlist](#)

### 5.6.2 멀티 워드 리스트 만들기

멀티 워드 리스트를 만드는 과정은 [여기](#)에 설명되어 있으며, 그 과정을 거치면 <그림 47>과 같은 결과가 나타난다.

N	Word	Freq.	%
1	SHE COULD NOT	54	0.06
2	SHE HAD BEEN	36	0.04
3	I AM SURE	35	0.04
4	A GREAT DEAL	34	0.04
5	HE HAD BEEN	34	0.04
6	IT WOULD BE	32	0.04
7	COULD NOT BE	29	0.03
8	I DO NOT	28	0.03
9	IT WAS A	28	0.03
10	AS SOON AS	23	0.03
11	HAD NOT BEEN	23	0.03
12	HE DID NOT	21	0.03
13	IN THE WORLD	21	0.03
14	THAT HE HAD	21	0.03
15	THE MISS MUSGROVES	21	0.03
16	THERE HAD BEEN	21	0.03
17	AND CAPTAIN WENTWORTH	20	0.02

<그림 47> 인덱스 파일 생성 결과

위와 같은 결과를 저장하기 위해 *Ctrl*/*F2*를 누르면, 워드스미스가 *\_index\_3-5-word clusters*와 같은 파일 이름을 제안한다. 참고로 이렇게 저장된 멀티 워드 리스트 파일은 나중에 일반적인 워드 리스트 파일처럼 열린다.

## 6. 키워드

### 6.1 개요

<그림 48>은 워드스미스로 키워드 리스트를 추출한 것이다.

The screenshot shows a Windows application window titled "Romeo And Juliet\_1st.kws". The menu bar includes File, Edit, View, Compute, Settings, Windows, and Help. The main area is a table with columns: N, Key word, Freq., %, RC. Freq., and RC. %. The table lists 11 entries:

N	Key word	Freq.	%	RC. Freq.	RC. %
1	ROMEO	115	0.48	115.00	0.01
2	TYBALT	47	0.19	47.00	
3	JULIET	40	0.17	45.00	
4	NURSE	39	0.16	83.00	0.01
5	LOVE	138	0.57	1,948.00	0.24
6	NIGHT	82	0.34	849.00	0.10
7	PARIS	27	0.11	81.00	
8	MONTAGUE	21	0.09	41.00	
9	THOU	277	1.15	5,358.00	0.66
10	O	161	0.67	2,615.00	0.32
11	MERCUTIO	14	0.06	14.00	

Below the table, there is a navigation bar with tabs: KWs (selected), plot, links, clusters, filenames, source text, and notes. At the bottom, it shows "32 entries" and "Row 0". A status bar at the bottom right says "loading J:\W".

<그림 48> 키워드 리스트

키워드는 연구 코퍼스를 레퍼런스 코퍼스와 비교했을 때 비정상적으로 자주 나타나는 단어들이다. <그림 48>을 자세히 살펴보면, 각 키워드 옆에 여러 숫자들이 있는데 그 숫자들은 소스 텍스트에서 특정 키워드가 얼마나 자주 사용되었는지 보여주며, 또한 레퍼런스 코퍼스에서는 해당 키워드가 얼마나 자주 사용되었는지를 보여준다.

<그림 48>은 로미오와 줄리엣과 셰익스피어 희곡 전체를 비교하여 추출한 로미오와 줄리엣의 키워드 리스트이다. 이 키워드 리스트를 통해 로미오와 줄리엣에 등장하는 주요 인물들의 이름과 thou 와 같은 대명사, love 와 night 와 같은 주제어가 로미오와 줄리엣의 주요 키워드인 것을 파악할 수 있다.

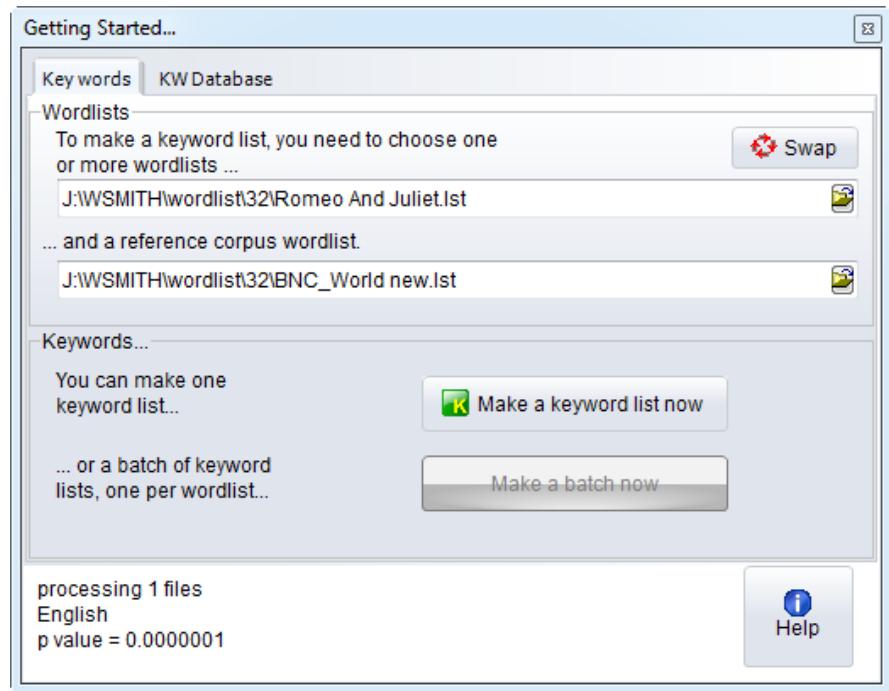
## 6.2 키워드 리스트 만들기

키워드 리스트를 만들려면, 우선 콘트롤러에서 키워드(KeyWords) 버튼을 클릭한다(<그림 49>)



<그림 49> 키워드 기능 선택하기

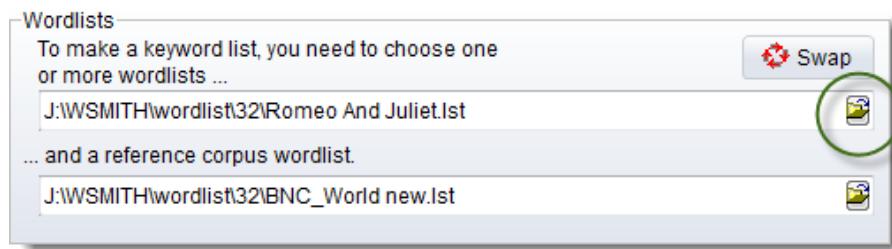
키워드 기능이 시작하면, 메뉴에서 *File*을 선택한 다음, *New*를 클릭한다. 그러면 <그림 50>과 같은 팝업창이 나타난다.



<그림 50> 키워드 리스트 추출하기

참고로 키워드 리스트를 추출하기 위해서는 미리 워드스미스로 해당 코퍼스의 워드 리스트를 만들어서 저장해 놓아야 한다.

<그림 51>에 표시된 것과 같이 풀더 모양의 아이콘을 클릭하면, 사용자가 원하는 워드리스트 파일을 찾을 수 있다.



<그림 51> 워드리스트 불러오기

참고로 연구 코퍼스의 키워드를 추출할 때, 비교 대상이 되는 레퍼런스 코퍼스의 규모가 크면 클수록 더 정확한 키워드 리스트를 추출하는데 도움이 된다.

연구 코퍼스의 워드 리스트와 레퍼런스 코퍼스의 워드 리스트를 선택하였으면, 이제 *Make a keyword list now* 버튼을 클릭한다. 이때 레퍼런스 코퍼스 워드 리스트를 불러오지 않으면, *Make a keyword list now* 버튼이 활성화되지 않으니, 만일 이 버튼이 활성화되지 않았다면 각 단계를 다시 확인해 본다.

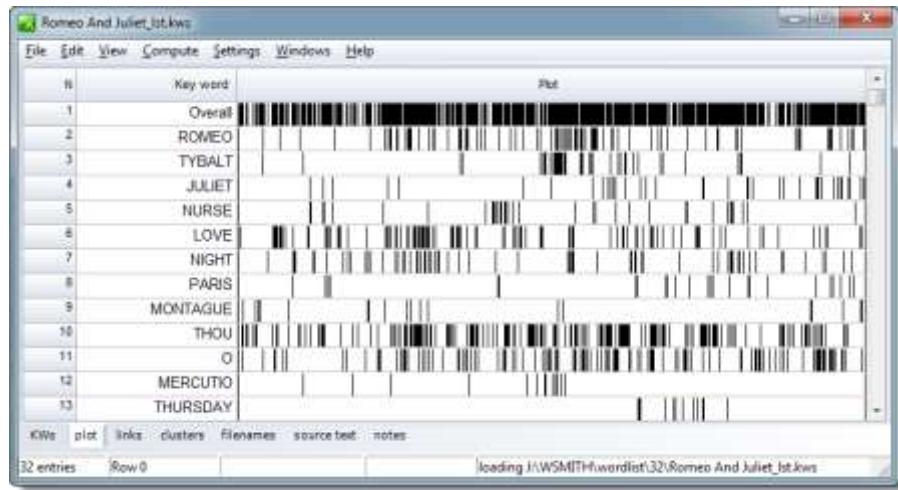
이상의 과정을 제대로 하였다면, <그림 52>와 같은 키워드 리스트가 추출된다.

N	Key word	Freq.	%	RC. Freq.	RC. %
1	ROMEO	115	0.48	115.00	0.01
2	TYBALT	47	0.19	47.00	
3	JULIET	40	0.17	45.00	
4	NURSE	39	0.16	83.00	0.01
5	LOVE	138	0.57	1,948.00	0.24
6	NIGHT	82	0.34	849.00	0.10
7	PARIS	27	0.11	81.00	
8	MONTAGUE	21	0.09	41.00	
9	THOU	277	1.15	5,358.00	0.66
10	O	161	0.67	2,615.00	0.32
11	MERCUTIO	14	0.06	14.00	

<그림 52> 로미오와 줄리엣의 키워드 리스트

### 6.3 키워드 플롯

<그림 53>은 셰익스피어의 모든 희곡과 비교한 로미오와 줄리엣의 키워드 플롯이다.

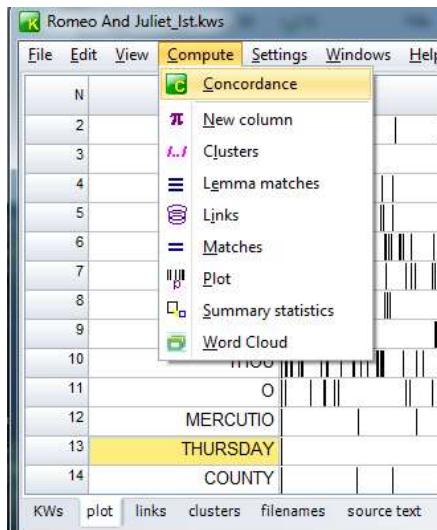


<그림 53> 로미오와 줄리엣의 키워드 플롯

<그림 53>에서 사용자는 각 키워드를 볼 수 있고, 각 키워드가 어느 위치에 출현하는지도 파악할 수 있다. 참고로 <그림 53> 키워드 플롯 막대의 왼쪽 끝은 텍스트의 시작을 의미하고, 오른쪽 끝은 텍스트의 끝을 의미한다. 키워드 플롯을 자세히 살펴보면, Love, Nurse 와 Tybalt 등이 로미오와 줄리엣의 다양한 지점에서 출현하는 것을 확인할 수 있다.

## 6.4 선택한 키워드 콘코던싱

키워드 리스트를 추출하면, 어떤 문맥에서 해당 키워드가 사용되었는지 볼 수 있다. 사용자가 관심있는 단어를 1개, 또는 그 이상 선택한다(<그림 54>).



<그림 54> 키워드 콘코던싱

메뉴에서 *Compute / Concordance* 를 선택한다. 참고로 여기서는 뭔가 흥미로울 것 같은 Thursday 를 선택하였다. 그러면 <그림 55>와 같은 결과를 얻을 수 있다.

The Concord software interface shows a list of 14 entries. The first entry is highlighted in yellow. The text in the list is as follows:

N	Concordance	%
1	else fail, myself have power to die. On Thursday, sir? the time is very short.	72%
2	to't, think on't, I do not use to jest. Thursday is near; lay hand on heart,	70%
3	I tell thee what, get thee to church o' Thursday, Or never after look me in the	70%
4	a wife. That may be must be, love, on Thursday next. What must be shall be.	73%
5	fit to furnish me to-morrow? No, not till Thursday; there is time enough. Go,	77%
6	must, and nothing may prorogue it, On Thursday next be married to this	74%
7	I should disturb devotion! Juliet, on Thursday early will I rouse you: Till	74%
8	And there an end. But what say you to Thursday? My lord, I would that	64%
9	is too soon; O' Thursday let it be: o' Thursday, tell her, She shall be	64%
10	ha! Well, Wednesday is too soon; O' Thursday let it be: o' Thursday, tell her,	64%
11	you to Thursday? My lord, I would that Thursday were to-morrow. Well, get	64%
12	But fettle your fine joints 'gainst Thursday next, To go with Paris to	69%
13	is that? Marry, my child, early next Thursday morn The gallant, young, and	68%
14	were to-morrow. Well, get you gone: o' Thursday be it then. Go you to Juliet	64%

At the bottom of the window, there are tabs for concordance, collocates, plot, patterns, clusters, timeline, breakdowns, and filenames. The concordance tab is selected. A status bar at the bottom shows '14 entries Row 4' and a search input field containing 'e what, get thee to church o' T'.

<그림 55> 키워드 Thursday 의 콘코던스 라인

### **<번역 후기>**

원본 영어 버전을 충실히 번역하다 보니, 매끄럽지 못한 표현이 보입니다. 또한, 원본 번역에 충실하다 보니 코퍼스 입문자가 이해하기에 어려운 내용과 중간중간에 다소 설명이 생략된 부분도 보입니다. 제한된 지면에서 원본의 스크린 캡쳐 이미지를 사용하다 보니 화질이 충분히 좋지 않습니다.

코퍼스연구소에서는 이러한 부분을 반영하여 우리나라 코퍼스 및 워드스미스 입문자 입장에서 손쉽게 따라할 수 있는 2 번째 튜토리얼과 동영상 데모를 준비하고 있습니다.

2 번째 버전을 기대해 주십시오.

To be updated.

문의: <http://icr.or.kr/contact>