ABM 연수 2일차 자료

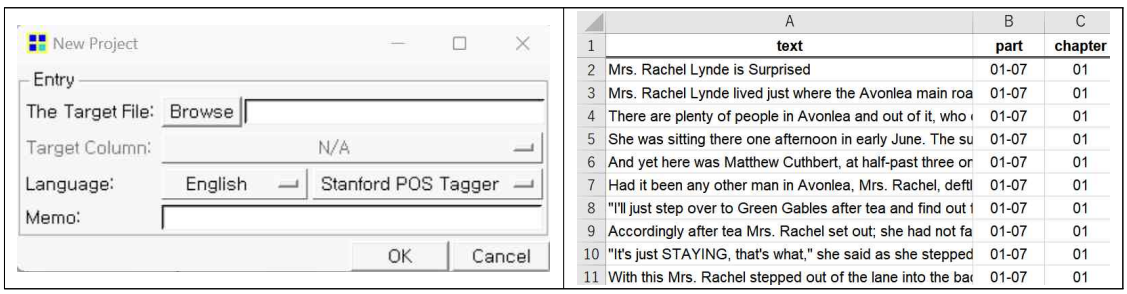
신동광(광주교대)
sdhera@gmail.com

1. KH Coder(텍스트 마이닝)

https://github.com/ko-ichi-h/khcoder/releases/download/3.Beta.06d/khcoder-3b06d.exe

1) 데이터 업로드 및 전처리

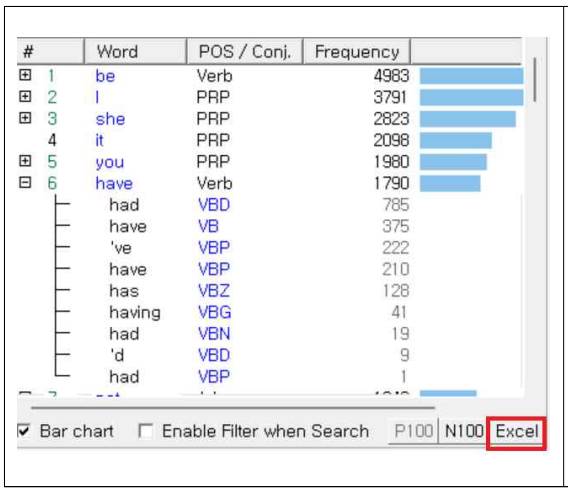
-Project > New > Browse(엑셀 데이터 불러오기) > English(분석언어 설정) > Pre-Processing > Run Pre-Preprocessing



	A	B	C
1	text	part	chapter
2	Mrs. Rachel Lynde is Surprised	01-07	01
3	Mrs. Rachel Lynde lived just where the Avonlea main roa	01-07	01
4	There are plenty of people in Avonlea and out of it, who	01-07	01
5	She was sitting there one afternoon in early June. The su	01-07	01
6	And yet here was Matthew Cuthbert, at half-past three or	01-07	01
7	Had it been any other man in Avonlea, Mrs. Rachel, deftl	01-07	01
8	"I'll just step over to Green Gables after tea and find out	01-07	01
9	Accordingly after tea Mrs. Rachel set out; she had not fa	01-07	01
10	"It's just STAYING, that's what," she said as she stepped	01-07	01
11	With this Mrs. Rachel stepped out of the lane into the bac	01-07	01

2) 품사가 구분된 레마(lemma) 목록 만들기

-Tools > Words > Frequency List > Term Freq(TF)/Document Freq(DF) 중 택 1 > Excel

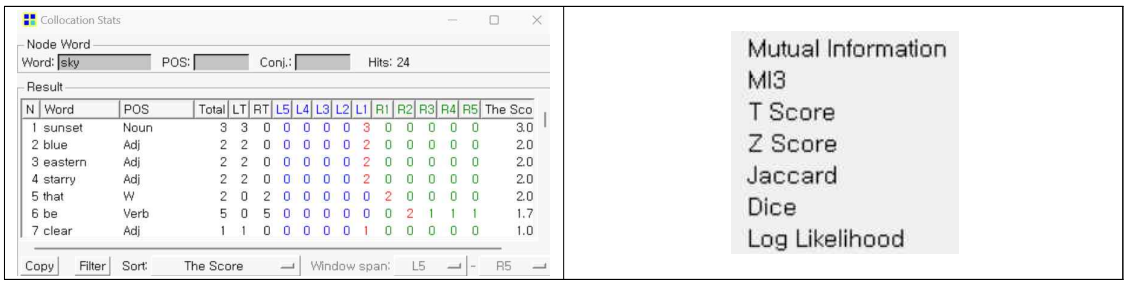


#	Word	POS / Conj.	Frequency
1	be	Verb	4983
2	I	PRP	3791
3	she	PRP	2823
4	it	PRP	2098
5	you	PRP	1980
6	have	Verb	1790
	had	VBD	785
	have	VB	375
	've	VBP	222
	have	VBP	210
	has	VBZ	128
	having	VBG	41
	had	VBN	19
	'd	VBD	9
	had	VBP	1

Noun		ProperNoun	
girl	267	ANNE	1138
thing	260	MARILLA	849
time	208	Diana	414
eye	152	Mrs	404
home	136	Matthew	361
child	134	Lynde	151
school	128	Miss	138
night	117	Mr	134
hair	114	Barry	132
hand	110	Gilbert	113

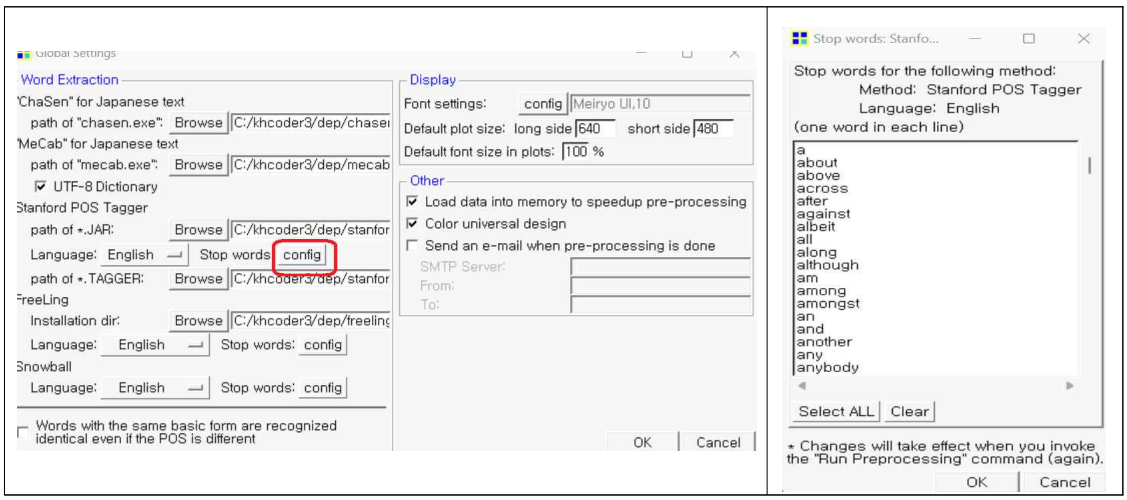
3) 연어 관계 분석하기

-Tools > Words > KWIC Concordance > 검색어 입력하기(예, sky) > Search > Stats > The Score로 연어 관계 분석 공식 변경 가능

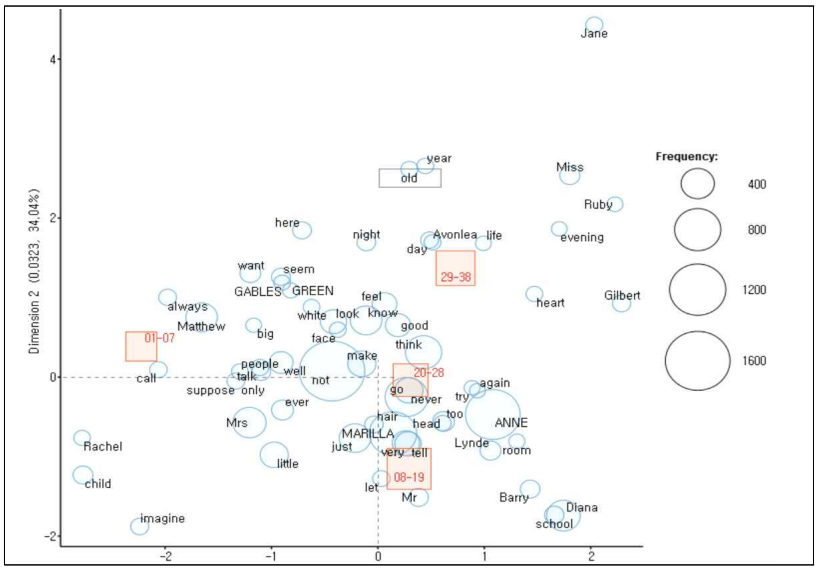


4) 이야기 흐름별 키워드 추출하기

-Project > Settings > Standford POS Tagger: Stop words: Config > Function word 입력 > Run Pre-Processing(내용어 분석만 원하면)



-Tools > Words > Correspondence Analysis > Words × Variable(s)에서 분석기준을 part/chapter 중 택 1(chapter가 너무 많으면 분석 결과의 해석이 어렵기때문에 큰 단위인 part를 권장), Bubble plot을 마우스 오른쪽 버튼으로 선택하면 주요 키워드 빈도수를 버블로 제시 > 특정 단어를 클릭하면 Concordance 예시 확인 가능

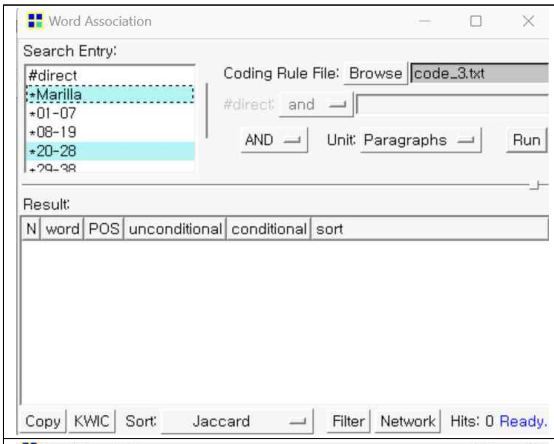


-각 Part를 나타내는 붉은색 박스를 기준으로 박스의 근처에 분포하는 어휘들이 각 part의 주요 키워드이며 세로축과 가로축 0 지점에 가까울수록 전체 이야기에 공통적인 키워드임

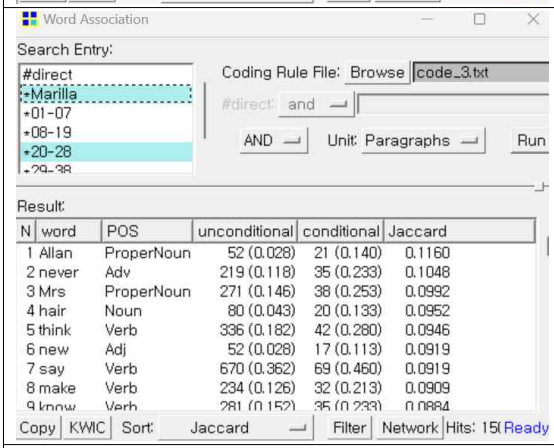
5) 특정 주제어와 연계된 각 part의 키워드 찾기

-Tools > Words > Word Association > Browse(Coding Rule File 불러오기, 예시 code_3)

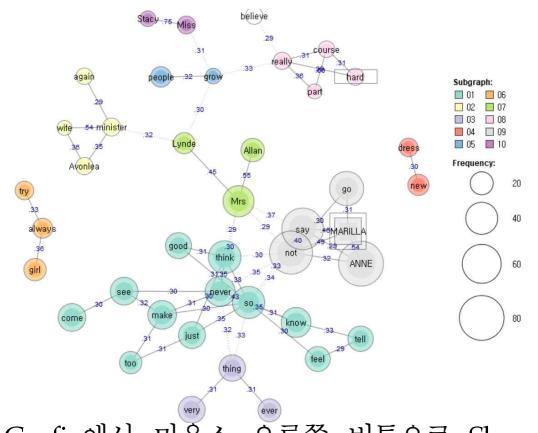
code_3 - 메모장 파일 편집 보기 *Marilla Marilla *01-07 <>part-->01-07 *08-19 <>part-->08-19 *20-28 <>part-->20-28 *29-38 <>part-->29-38	-*는 산출 결과의 제목을 의미하며 그 아래는 분석 대상(예, Marilla) -각 part가 포함하는 chapter의 범위 기술

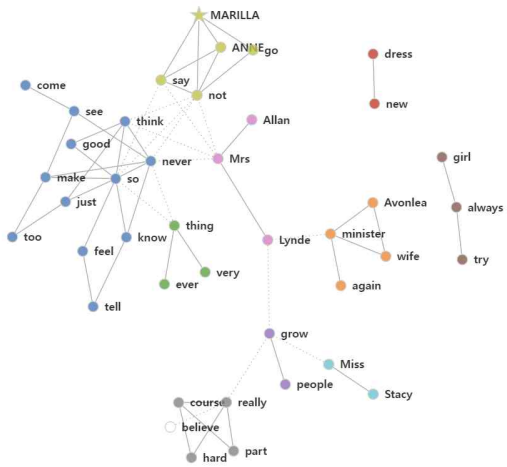
<table>
<tr>
<td> 
</td>
<td>-분석 키워드인 Marilla를 클릭하고 이 키워드와 연계어를 분석하기 위해 분석하고자 하는 part를 선택, part를 선택할 때는 Ctrl 키를 누르고 마우스로 원하는 part 선택

-Unit에서 분석단위를 Sentence로 할 것인지 Paragraph로 할 것인지 택 1

-분석을 위해 Run 클릭</td>
</tr>
<tr>
<td> 
</td>
<td>-part 3(chapters 20-28)에서 Marilla와 가장 밀접한 단어는 “Allan”이란 캐릭터

-그래픽으로 확인하기 위해서는 Network 버튼 클릭

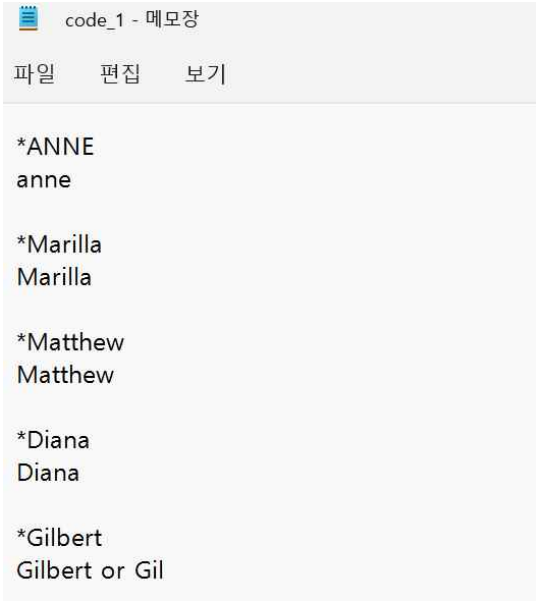
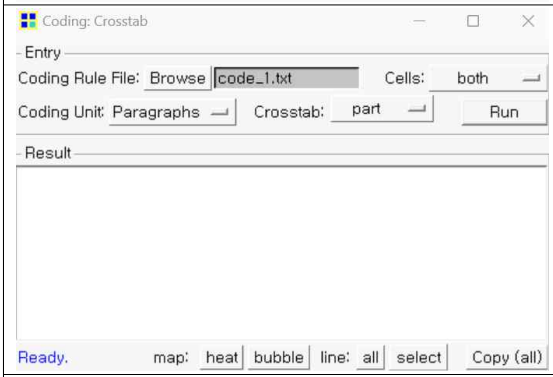
-Jaccard는 어휘 간의 거리를 측정하는 알고리듬이지만 절대빈도 등 여러 공식 선택 가능, 공식에 따라 수치와 그래픽이 달라짐</td>
</tr>
<tr>
<td> 

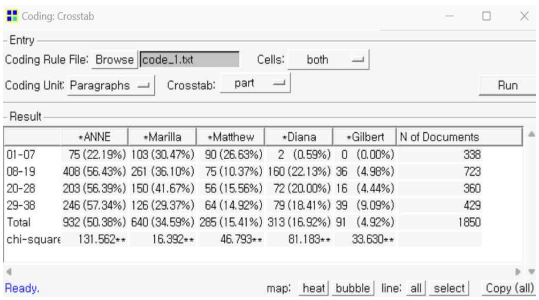
Config에서 마우스 오른쪽 버튼으로 Show Coef.를 선택하면 계수와 함께 제시</td>
<td> 

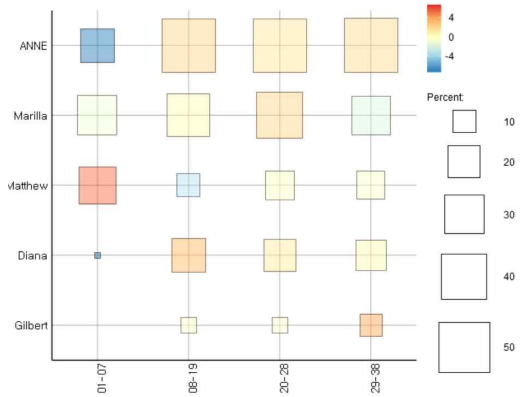
HTML 버전</td>
</tr>
</table>

6) 여러 주제어들 간의 part/chapter별 평균빈도(χ^2) 비교하기

-Tools > Coding > Crosstab > Browse(Coding Rule File 불러오기, 예시 code_1)

<table>
<tr>
<td> 
</td>
<td>-5명의 주요 등장 인물 평균빈도 비교</td>
</tr>
<tr>
<td> 
</td>
<td>-Coding Unit에서 빈도분석 단위를 Sentence 또는 Paragraph 택 1

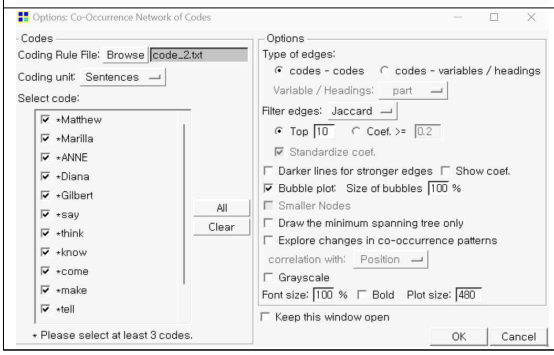
-Crosstab에서 part에서 빈도를 분석할지 chapter에서의 빈도를 분석할지 택 1</td>
</tr>
<tr>
<td> 

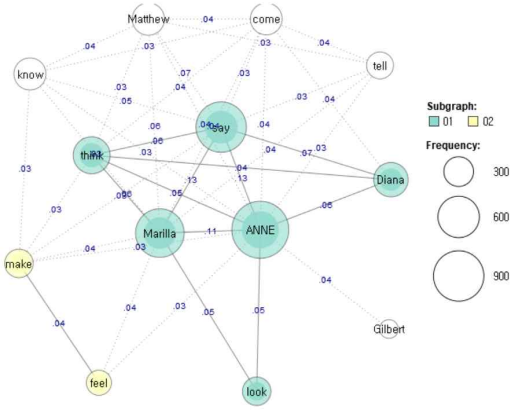
각 등장 인물의 part별 빈도비율이 제시되어 어떤 part에서 비중이 큰지를 확인</td>
<td> 

bubble을 선택하면 그래픽으로 각 part의 중요 인물 확인 가능</td>
</tr>
</table>

6) 특정 주제어와 특정 연계어 간의 관계성 분석하기(예, 등장 인물들과 감정동사 간의 관계 분석)

-Tools > Coding > Co-Occurrence Network > Browse(Coding Rule File 불러오기, 예시 code_2)

*Matthew Matthew *Marilla Marilla *ANNE anne *Diana Diana *Gilbert Gilbert ##################################### #*say #say *think think *know know #*come #come	-code_2에서 분석 어휘 앞에 #있는 경우는 메모처럼 사용하기 위해 입력하기 때문에 메모가 필요없으면 사용하지 않음
	-Coding Unit에서 빈도분석 단위를 Sentence 또는 Paragraph 택 1 -수치가 잘 나오지 않으면 Top 10을 조정 > Top 40 등

	-Marilla와 Anne 사이에 say 동사가 가장 많이 연결되어 두 사람 사이의 대화가 많다는 것을 알 수 있고 Anne과 Diana, 그리고 Anne과 Gilbert의 관계성은 보이지만 Diana와 Gilbert의 직접적인 관계성은 없다는 것을 확인 가능

2. Familizer25(어휘목록 제작)

https://cafe.daum.net/sdhera/G3Jv/132

어휘족 목록을 개발하기 위해서는 먼저 Nation과 Webb(2011)의 BNC-COCA 25,000에 포함된 1-25까지의 basewrds를 basewrd1 하나로 모두 복사 또는 잘라내어 합친다. 이것으로 소위 Familizer25라 한다. 만약 하이픈으로 연결된 합성어까지 포함하고 싶다면 1-28까지의 basewrds를 통합한다. 이 프로그램을 Familizer28이라 한다. 통합된 파일은 일종의 어휘족 은행 또는 데이터베이스로 활용된다.

(1) 그 다음은 어휘족의 대표형을 추출할 소스로 사용할 코퍼스(텍스트)를 선택해야 한다. 사용목적에 따라서 코퍼스는 교과서가 될 수도 있고 특정 분야의 언어자료가 될 수 있다. 예를 들어 10개의 코퍼스를 소스로 활용한다고 가정할 때 보다 일반적인 단어를 추출하기 위해서는 구어자료(spoken corpus)의 비율을 높여야 한다. 반면 문어자료(written corpus)의 비율이 높이면 보다 어려운 단어들이 포함될 수 있다.

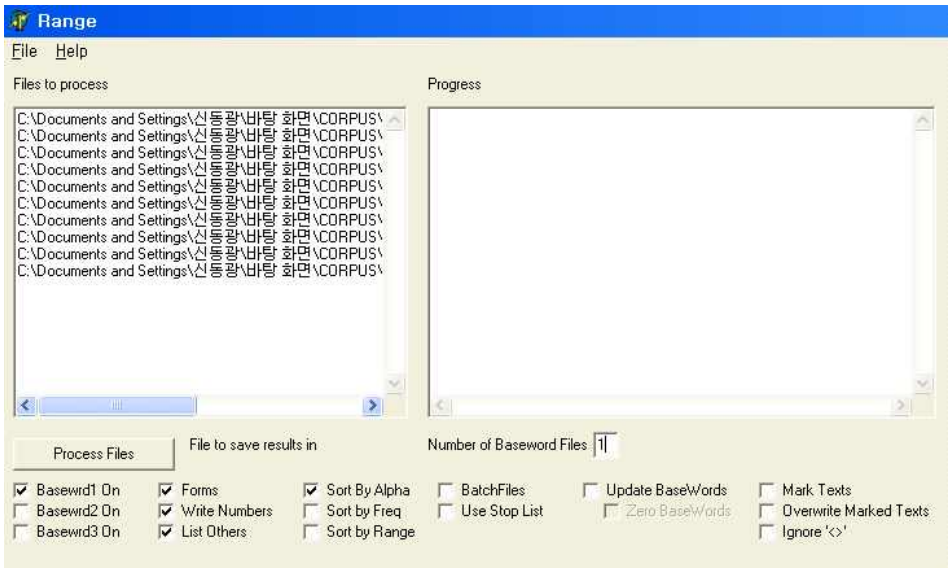
(2) 단어 선별의 기준은 위에서 언급한 바와 같이 빈도수(frequency), 사용범위(range), 사용분포(dispersion)를 고려해 볼 수 있다. 소스선택과 일관성을 유지하려면 보다 일반적으로 사용되는 단어를 선별해야 하고 그러기 위해서는 사용범위를 가장 우선순위에 두어야 한다. 사용분포는 단어형태(type)에만 적용할 수 있는 관계로 Familizer만을 이용하여 어휘족 목록을 제작하고자 한다면 기준에서 제외 될 수 밖에 없다 또한 빈도수와 사용분포는 상관관계가 상당히 높아 제외한다 하여도 결과 값에는 큰 영향을 미치지 않는다. 빈도수와 사용범위의 수치를 구하기 위해 다음과 같은 절차를 따른다.

(2-1) [그림 15] 왼쪽 상단에 보이는 File을 클릭하고 Open을 클릭하여 분석하고자 하는 코퍼스들을 불러온다.

(2-2) 먼저 분석한 결과 값을 Save를 클릭하여 원하는 이름의 파일로 저장한다. 일반적으로 여타 프로그램에서는 결과 값은 분석 후에 저장하도록 되어 있지만 본

프로그램에서는 저장 후 분석을 진행한다는 점이 색다르다.

(2-3) 사용하는 어휘족 목록이 basewrd1으로 통합되어 있는 관계로 하단의 Number of Baseword Files에는 1로 입력하고 Basewrd1 On만 선택한다. Basewrd2 On과 Basewrd3 On에는 클릭하여 선택을 해제한다. 분석 어휘의 정렬 순서를 위해 Sort by Freq(빈도순) 또는 Sort by Range(사용범위순)를 선택한다.



[그림 15] 프로그램 실행과정

(3) [그림 16]와 같이 결과분석 파일을 열고 LIST OF FAMILY GROUPS를 찾는다.

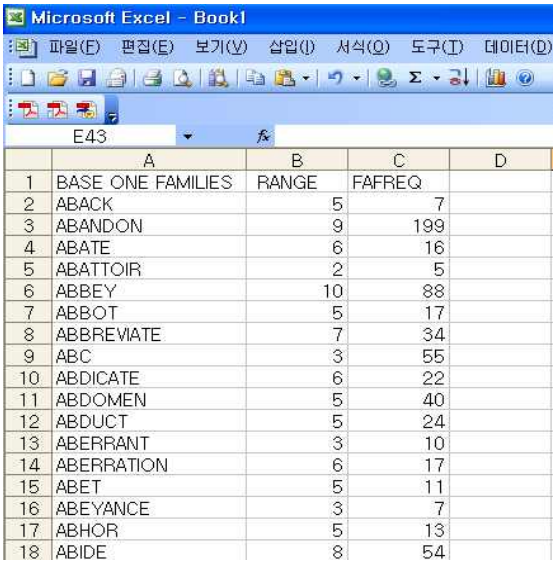
LIST OF FAMILY GROUPS

BASE ONE FAMILIES	RANGE	TYFREQ	FAFREQ	F1	F2	F3	F
ABBEY	10	86	88	4	1	12	
ABOVE	10	1166	1166	48	41	31	8
ABROAD	10	215	215	19	17	12	2
ABRUPT	10	42	86	10	1	1	
ABSENCE	10	221	240	8	2	3	4
ABUSE	10	251	404	3	11	5	2
ACADEMY	10	75	306	4	7	3	1
ACCELERATE	10	24	234	6	18	1	
ACCENT	10	162	232	55	19	41	
ACCEPT	10	935	2197	38	57	50	6
ACCESS	10	595	695	22	20	6	1
ACCIDENT	10	452	691	46	54	35	1
ACCOMMODATE	10	122	436	18	9	6	1
ACCOMPANY	10	38	240	1	1	2	1
ACCORDING	10	684	684	32	31	22	3
ACCORDINGLY	10	114	114	1	3	1	1

[그림 16] 결과분석

위의 자료를 복사해서 엑셀(EXCEL) 프로그램에 붙여 넣는다.

(3-1) [그림 17]과 같이 엑셀에 자료를 붙여 넣은 후 BASE ONE FAMILIES, RANGE, FAFREQ의 값만 남기고 나머지는 삭제한다.


Microsoft Excel - Book1

파일(F) 편집(E) 보기(V) 삽입(I) 서식(O) 도구(T) 데이터(D)

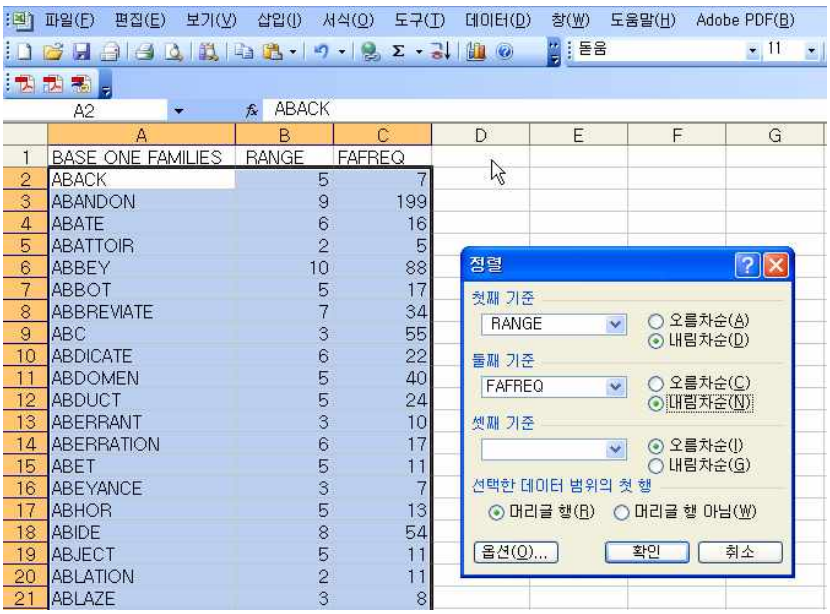
E43

	A	B	C	D
1	BASE ONE FAMILIES	RANGE	FAFREQ	
2	ABACK	5	7	
3	ABANDON	9	199	
4	ABATE	6	16	
5	ABATTOIR	2	5	
6	ABBEY	10	88	
7	ABBOT	5	17	
8	ABBREVIATE	7	34	
9	ABC	3	55	
10	ABDICATE	6	22	
11	ABDOMEN	5	40	
12	ABDUCT	5	24	
13	ABERRANT	3	10	
14	ABERRATION	6	17	
15	ABET	5	11	
16	ABEYANCE	3	7	
17	ABHOR	5	13	
18	ABIDE	8	54	

[그림 17] 엑셀 데이터 입력

(3-2) 엑셀의 자료에서 정렬부분을 지정하고 데이터 정렬 기능에서 첫째기준과 둘째 기준에

사용범위(Range)든 빈도수(Frequency)든 사용목적이나 사용자의 선택에 따라 입력하고 내림차순으로 [그림 18]와 같이 재정렬을 시행한다.



[그림 18] 엑셀 데이터 정렬

등급별 어휘목록을 개발하기 위해 개별 어휘의 순위를 결정하는 방식으로는 위에서 살펴본 우선순위 선정 기준의 순차적용 방식이 지금까지는 가장 일반적이었지

만 그 방식에는 몇 가지 기준을 추가하여 다음의 <표 9>와 같이 다양한 적용이 가능하다(신동광, 2018).

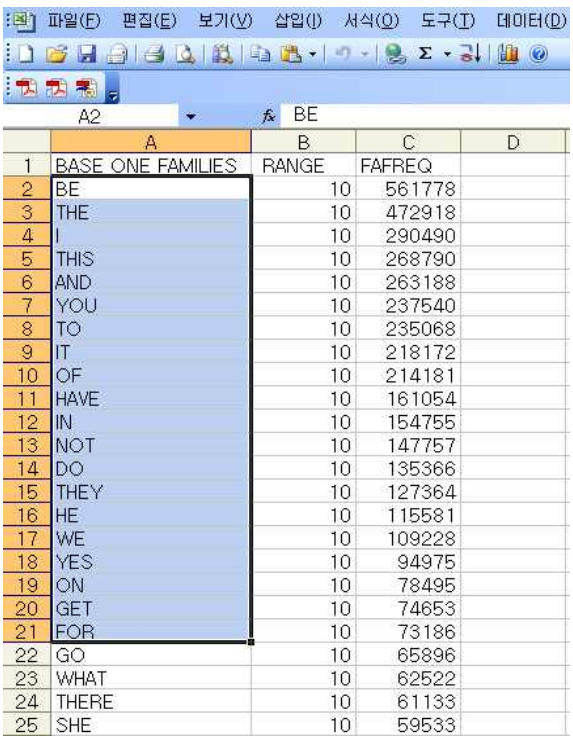
<표 9> 어휘 등급 구분을 위한 순위 확정 방식 예시

순위 확정 방식	기준 적용 절차	비고
유형 1	친숙도>초등권장 어휘 추천>사용범위>빈도수 순차 적용 순위	
유형 2	t점수 변환 후 가중치를 곱한 값들의 합을 기준으로 한 순위	친숙도×0.4+초등권장 어휘 추천×0.3+사용범위×0.2+빈도수×0.1
유형 3	t점수 변환 후 가중치를 곱하지 않은 값들의 합을 기준으로 한 순위	
유형 4	원점수에 가중치를 곱한 값들의 합을 기준으로 한 순위	친숙도×0.4+초등권장 어휘 추천×0.3+빈도수×0.2+사용범위×0.1
유형 5	원점수에 가중치를 곱하지 않은 값들의 합을 기준으로 한 순위	친숙도+초등권장 어휘 추천+사용범위+빈도수
유형 6	빈도수>사용범위>초등권장 어휘 추천>친숙도 순차 적용 순위	
유형 7	t점수 변환 후 가중치를 곱한 값들의 합을 기준으로 한 순위	빈도수×0.4+사용범위×0.3+초등권장 어휘 추천×0.2+친숙도×0.1
유형 8	원점수에 가중치를 곱한 값들의 합을 기준으로 한 순위	빈도수×0.4+사용범위×0.3+초등권장 어휘 추천×0.2+친숙도×0.1

위의 <표 9>에 제시된 8개의 선정 방식은 어휘 선정 기준을 순차적으로 적용하는 방식과 합산하는 방식, 원점수를 적용하는 방식과 t점수로 치환하여 적용하는 방식, 각 어휘 선정 기준을 동일한 비율로 적용하는 방식과 0.1배에서 0.4배까지 가중치를 두고 적용하는 방식으로 구분된다. 만약 기존의 방식과 마찬가지로 유형 1과 같이 우선순위 선정 기준의 순차적용 방식을 적용하게 되면 첫 번째 기준이 순위 확정에 거의 절대적인 영향을 주기 때문에 다양한 기준을 균형있게 반영하기 어렵다(1순위 값으로 순위를 나열한 후 동일한 값을 가지는 항목에 2순위의 값을 적용하여 재정렬하고 2순위의 값도 같을 경우 3순위의 값을 적용하는 방식). 이를 보완하기 위한 방식이 각 기준의 값을 통합하여 하나의 수치로 순위를 확정하는 방식이 제안된 것이다. 하지만 각 기준의 수치 단위가 다를 경우 수치의 단위가 큰 기준이 절대적인 영향을 주기 때문에 수치의 단위를 같게 조정할 필요가 있다(위에서 제시된 t값을 의미). 하지만 통합값을 적용할 경우에도 어휘목록의 개발의 목적을 고려하여 특정 기준에 가중을 적용할 필요도 제기될 수 있고 이를 반영한 방식이 유형 2라고 볼 수 있다.

(3-3) [그림 19]와 같이 선별 기준에 따라 표제어(대표형) 정렬이 끝나면 예외항

목의 여부를 결정해야 한다. 고유명사나 감탄사는 예외항목으로 두는 것이 일반적이며 이를 위해서는 표제어에서는 삭제를 해야 한다. 알파벳, 숫자, 나라명, 단위명칭, 외래어 등은 예외항목으로 제외시킬 것인지, 어휘족 목록에 포함시킬 것인지를 선택하고 예외항목으로 두고자 한다면 표제어 목록에서는 고유명사나 감탄사와 마찬가지로 삭제해야 한다. 그러한 선별과정이 끝나면 어휘족 목록의 단위를 결정하여 그 단위만큼 표제어를 복사하여야 한다. 즉 수준별 단어목록을 제작한다고 할 때 한 수준이 올라갈 때 마다 추가되는 어휘족의 양을 100단어, 250단어, 500단어, 1,000단어 단위 등으로 결정해야 한다. 1,000단어 단위로 결정을 하게 되면 수준별로 각 1,000단어씩 증대되게 된다. [그림 19]는 20단어씩 증대되는 어휘족 목록을 제작하는 예시로 첫 20단어의 표제어를 복사하여 어휘족을 목록을 만들고 그 다음으로 그 다음 20단어를 복사하여 어휘족 목록을 만들면 된다.



	A	B	C	D
1	BASE ONE FAMILIES	RANGE	FAFREQ	
2	BE	10	561778	
3	THE	10	472918	
4	I	10	290490	
5	THIS	10	268790	
6	AND	10	263188	
7	YOU	10	237540	
8	TO	10	235068	
9	IT	10	218172	
10	OF	10	214181	
11	HAVE	10	161054	
12	IN	10	154755	
13	NOT	10	147757	
14	DO	10	135366	
15	THEY	10	127364	
16	HE	10	115581	
17	WE	10	109228	
18	YES	10	94975	
19	ON	10	78495	
20	GET	10	74653	
21	FOR	10	73186	
22	GO	10	65896	
23	WHAT	10	62522	
24	THERE	10	61133	
25	SHE	10	59533	

[그림 19] 어휘족(family) 목록 제작을 위한 표제어(headword) 선택

(4) 어휘족 목록을 만들기 위해서는 복사한 표제어를 메모장(텍스트 파일)에 붙여 저장하고 Familizer의 open을 클릭하여 저장한 파일을 불러오고 basewrd1 on만 선택, Number of Baseword Files를 1로 수정, Update basewords, Zero basewords를 클릭한다. 그 다음 Save한 후 Process Files를 클릭하면 Range 값이 저장된다. 실제로는 분석 값은 불필요한 것이지만 절차 상 필요한 과정이다.

(5) Familizer를 통해 분석을 마치고나면 [그림 20]과 같이 상단의 File를 클릭

하여 Move/Copy Basewords를 선택하고 greater than 0을 클릭, Copy words를 클릭, basewrd4를 본인이 원하는 이름으로 입력, 단 basewrd1을 제외한 다른 이름 또는 그대로 basewrd4로 써도 상관없다.

[그림 20] 선택 어휘족(family) 복사

(6) Move/Copy Basewords 기능을 통해 [그림 20]에서 OK를 클릭하면 지정한 이름, 즉 basewrd4의 이름으로 다음과 같은 어휘족 목록(family list)이 생성된다. 단, Nation의 14,000단어 내에 포함된 어휘족만 생성된다. basewrd4를 basewrd1로 바꿔 따로 저장하고 [그림 19]에서 분석한 그 다음의 단어를 선택하여 똑같은 절차로 어휘족을 생성하여 basewrd2로 저장하고 계속 원하는 만큼 반복하면 사용자만의 수준별 어휘족 목록이 작성 된다. RANGE 프로그램 폴더에 기존의 basewrd를 삭제하고 제작한 어휘족 목록을 붙여 넣으면 새로운 RANGE 프로그램의 제작이 끝난다.

[그림 21] 어휘족(family) 목록 생성

(7) 고유명사와 같은 예외항목이 있을 경우 basewrd를 하나 더 만들고 그 안에 [그림 22]과 같이 추가 항목을 입력해 주면 된다.

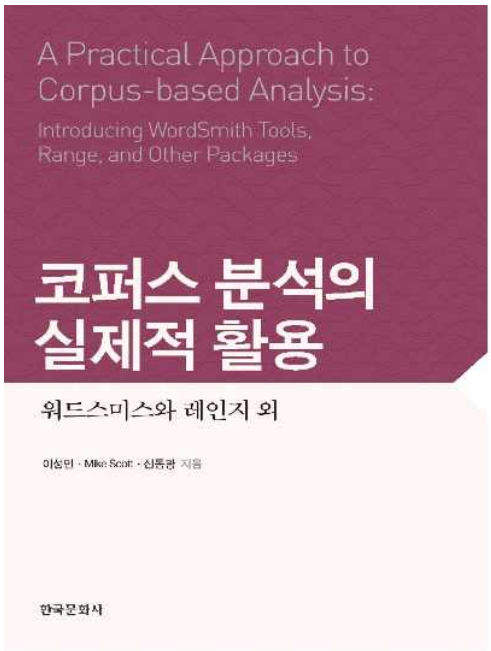
```
kimchi 0
    gimchi 0
america 0
    american 0
    americans 0
minsu 0
minhee 0
```

[그림 22] 어휘족(family) 목록 생성

※연어 분석/제작 도구인 COCA_MWU_ColloGram은 프로그램 내의 매뉴얼 참고

참고문헌

이성민, Mike Scott, 신동광. (2018). *코퍼스 분석의 실제적 활용: 워드스미스와 레인지 외*. 한국문화사.



참고: https://cafe.daum.net/sdhera