
COCA_MWU20 ColloGram 사용자 설명서(2018)

개발자:

신동광(광주교육대학교)
전유아(한양대학교)
이신웅(한양대학교)
박명수(상명대학교)

1. 프로그램 개요

기존 언어 분석 프로그램은 특정 언어 목록을 기준으로 분석하기 보다는 키워드(node)를 입력하고 좌우에 나타나는 언어소(collocate)를 검색하거나 반복되는 ‘N-gram’의 패턴을 주로 분석하는 프로그램이었다. 하지만 ColloGram은 특정 언어 목록을 탑재하여 분석 대상 자료에 포함된 언어 사용을 분석하는 프로그램이다. ColloGram은 ‘Collocation’과 ‘N-gram’ 또는 ‘Program’의 합성어로 언어 분석 프로그램을 의미하며 프로그램에서 지원하는 기능은 어휘 분석 프로그램인 Heatley와 Nation(2002)의 RANGE program에서 지원하는 기능과 유사하게 개발되었다. 프로그램에 탑재된 언어¹⁾ 목록은 1990년부터 2015년까지 미국의 언어 데이터로 구축된 5억 단어의 Corpus of Contemporary American English(COCA) 중 일반 연구자에게 COCA가 판매 되던 시점인 2014년도를 기준으로 1990년에서 2009년까지의 4억 5천만 단어의 데이터에 기반하여 추출되었고 추출 기준으로는 약 4억 5천만 단어 데이터에서 최소 20회 이상의 빈도를 가지고 있으며 하나의 독립된 의미 단위를 구성한다는 조건을 적용하였다. 또한 아래 <표 1>에 제시된 COCA의 5개 대영역 가운데 최소 4개 영역에서 출현하는 조건, 즉 사용범위(range) 4 이상을 적용하였다.

<표 1> COCA의 대영역 분류 및 규모

Academic	Fiction	Magazine	Newspaper	Spoken	총계
87,600,712	85,496,648	92,292,104	88,503,944	94,959,712	448,853,120

또한 학계에서는 처음으로 다어휘군(Multi-Word Unit family)이라는 개념을 도입하였다. 다어휘군(MWU family)은 어휘군(word family)과 유사한 개념으로 대표형을 기준으로 하여 굴절과 파생 변화형을 모두 포함하는 광의적인 개념이다. 다어휘 표현의 목록을 제작하다 보면 예를 들어 “this moment”와 “at this moment”를 별개의 다어휘 유형으로 취급해야 하는 지에 대한 고민에 빠질 때가 있다. 이 두 개의 다어휘 유형을 개별 검색한다고 가정해 보면 “this moment”가 30회, “at this moment”가 20회가 나올 수 있다. 하지만 “this moment”의 검색에는 “at this moment”의 한 부분인 “this moment”를 총 빈도에 포함하기 때문에 순수한 “this moment”의 빈도를 산출하기 위해서는 “this moment”의 빈도에서 “at this moment”의 빈도를 제외할 필요가 있다. 이와 같이 본 연구에서의 다어휘군의 개념은 빈도 산출의 중복을 없애면서 대표형에 어휘가 추가되거나 삭제되는 표현을 다어휘군의 파생형으로 간주하고 동사의 경우 동사의 굴절 변화형(예, go home, goes home, going home, went home, gone home)과 명사의 경우 단복수(예, year old,

1) ColloGram의 설정에서는 ‘collocation(언어)’이란 용어로 제시되며 이는 다어휘 표현(Multi-Word Unit, MWU)과 동일한 개념을 사용됨

years old)를 포함하여 이를 다어휘군의 굴절형으로 정의하였다. 결과적으로 빈도수와 사용범위의 기준을 충족하는 COCA 다어휘군 목록은 굴절형과 파생형을 포함하는 대표형 기준으로 10,214개가 선정되었고 이중 상위 10,000개(굴절형과 파생형을 구분한 다어휘 유형은 총 31,680개)를 최종 선별하여 사용범위를 1순위, 빈도수를 2순위로 순위를 확정하고 500개를 한 개의 등급으로 구분하였다. 이를 바탕으로 ColloGram에는 총 20개 등급화된 다어휘군 목록이 탑재되었다. 그리고 이 프로그램과 다어휘군 목록을 통칭하여 *COCA_MWU20*이라 한다.

끝으로 ColloGram은 연속된 총 2-10개의 단어로 구성된 다어휘군 유형을 분석할 수 있으며 백만 단어의 코퍼스 분석에 최적화 되어 있으며(분석 시간 10초 이내) 1회 최대 1억 단어까지 분석이 가능하나 이때는 분석 시간이 길게는 20여분까지 소요될 수 있다.

1.1. 프로그램 구성

- ① 단어회군 출현형(token), 단어회 유형(type), 단어회군(family) 빈도 분석
- ② 단어회군 기본형(head collocation)을 기준으로 단어회군의 구성 유형(단어회군의 기본형에 포함되는 파생, 굴절 변이형) 복사
- ③ 목록 중 기본형 목록(head collocation list) 추출
- ④ 목록 내 중복 제거

1.2. 프로그램 구성 파일

프로그램은 기본 제공되는 단어회군 목록(Basecollo1.txt, Basecollo2.txt, Basecollo3.txt...Basecollo20)과 실행파일(32bit, 64bit 용)로 구성되며 별도의 설치 과정 없이 바로 사용이 가능하다.

- ① Basecollo1.txt~Basecollo20.txt: 1-20까지는 사용범위 > 빈도순으로 각 500개 단어회군 포함
- ② Collogram32.exe: 일반적인 Desktop용
- ③ Collogram64.exe: 최신 Laptop용

1.2.1. 단어회군 목록 파일 구조

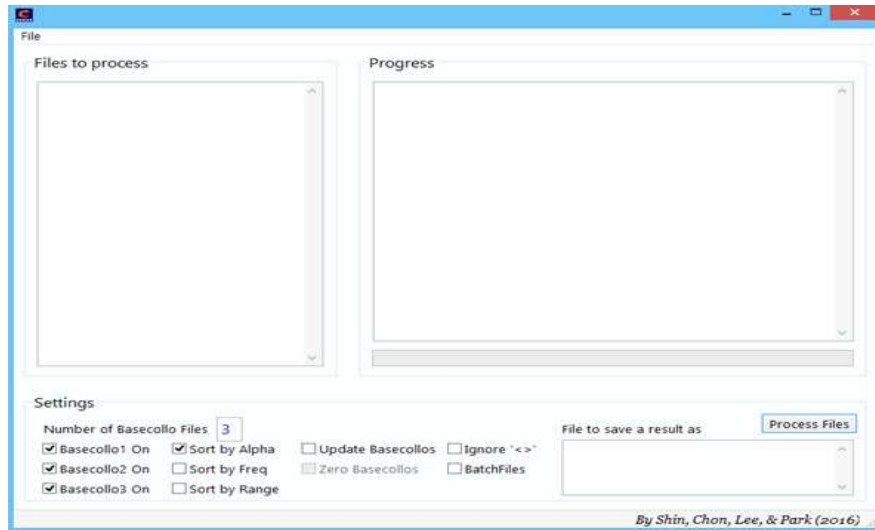
단어회군 목록의 구조는 “총 빈도수(모든 단어회군 유형의 빈도 총합)+탭(Tab)+단어회군 기본형(head MWU)+스페이스(space) 1칸+기본형의 빈도수”로 구성되어 있다. 그 아래의 단어회군 변이형은 탭을 한 번 더 들여 쓴 후 “단어회군 하위 유형 + 각 유형의 빈도수”로 구성된다. 총 빈도수나 타입의 빈도수 정보는 입력하지 않아도 프로그램은 실행된다. 다시 말해, 단어회군의 기본형 아래는 기본형에서 굴절된 변이형(동사의 경우, 굴절 변화형, 명사의 경우 단/복수)과 파생된 변이형(기본형에서 어순이 바뀌거나 단어회군 구성소가 추가되거나 축소된 변화형) 모두 열거한다. 현재 Basecollo에 포함된 단어회군은 기본형을 포함한 모든 하위 유형이 COCA에서 최소 빈도수 20과 사용범위 4 이상의 기준을 만족시키고 있다.

예) 아래 세 가지 형태의 ‘basecollo’ 목록 탑재 가능

빈도수 정보를 포함한 경우(단어회군 목록 기본 포맷)	14544	a lot more	13230
		a lot more to	707
		a whole lot more	607
	29207	a lot of people	27206
		lots of people	2001
	8186	all of a sudden	7076
		then all of a sudden	1110
	32776	all over	19912
		all over again	2374

	all over the country 3196 all over the world 4789 all over the place 2505	
빈도수 정보를 제외한 경우	a lot more a lot more to a whole lot more a lot of people lots of people all of a sudden then all of a sudden all over all over again all over the country all over the world all over the place	
하위 유형이 없는 경우(각 유형의 빈도수 및 사용범위 분석에 활용)	a lot more a lot of people all of a sudden all over	a lot more a lot more to a whole lot more a lot of people lots of people all of a sudden then all of a sudden all over all over again all over the country all over the world all over the place

1.3. 다어휘군 분석 과정



[그림 1] 다어휘군 분석 화면

1.3.1. 분석 파일 지정

상단의 File-Open 메뉴를 이용하여 분석하고자 하는 파일을 선택한다. 파일 선택은 하나 이상의 파일을 지정할 수 있다. 또는 drag & drop 기능이 있어 마우스로 분석하고자 하는 파일을 끌어와 박스 안에 넣으면 자동으로 분석 파일이 지정된다.

1.3.2. 저장 파일 지정

File-Save 메뉴를 이용하여 분석 후 저장될 파일의 지정한다. 만약 저장 파일명을 지정하지 않고, Settings의 BatchFiles를 선택하고 Process Files 버튼을 클릭하면 자동으로 분석 파일에 대한 결과가 “분석파일명_collo.txt”라는 이름으로 분석 파일이 있던 폴더에 저장된다.

1.3.3. 분석 옵션 지정

Number of Basecollo Files:

분석에 적용될 다어휘군 목록의 숫자를 지정한다. 예를 들어 위의 그림 1에 명시된 Basecollo의 수는 세 개에 불과하지만 프로그램 폴더 안에 더 많은 다어휘군 목록이 탑재되어 있을 경우 분석에 적용하고 싶은 다어휘군 목록의 개수를 새로 입력할 수 있다. 현재는 Basecollo1~Basecollo20까지 총 20개의 다어휘군 목록을 선택할 수 있다. 만약 새로운 Basecollo 파일을 개발하여 탑재하고 이중 앞의 3개 다어휘군 목록에 예외 항목을 입력하여 분석에서 때에 따라 제외하고 싶다면 Basecollo1 On, Basecollo2 On, Basecollo3 On을 다시 클릭하여 선택에서 제외하면 된다.

Sort by Alpha:

분석 결과의 단어회군 목록을 알파벳순으로 정렬하고자 하는 경우 사용한다.

Sort by Freq:

분석 결과의 단어회군 목록을 총 빈도수 순으로 정렬하고자 하는 경우 사용한다.

Sort by Range:

분석 결과의 단어회군 목록을 사용범위(얼마나 다양한 텍스트에서 사용되었는가를 측정) 순으로 정렬하고자 하는 경우 사용된다.

Update Basecollos:

현재 Basecollo는 2009년까지의 COCA 데이터를 기반으로 각 단어회군 유형의 빈도와 하나의 기본형이 포함하는 모든 하위 유형의 총 빈도수를 포함하고 있는데 새로운 코퍼스를 분석 시 Update Basecollos 선택하여 분석하면 기존 빈도수에 새로 분석한 코퍼스의 빈도수가 합산(add-up)되어 저장된다.

Zero Basecollos:

Update Basecollos 적용 시 Zero Basecollos 를 추가 선택하고 분석하면 기존 Basecollo에 포함된 COCA의 단어회군 빈도 정보가 사라지고 새로 분석한 코퍼스 파일 내의 단어회군 빈도수로 대체된다. 새로 분석한 파일에 기존의 단어회군 유형이 포함되어 있지 않은 경우는 빈도는 0으로 표시된다.

Ignore ‘<>’:

분석 파일에서 “<”로 시작하는 부분부터 “>”로 마치는 부분까지를 분석에서 제외할 때 사용한다. 보통 코퍼스에서< >을 이용하여 태그(Tag)를 달 때 분석에서 태그를 제외하고 원문만을 분석할 수 있다.

BatchFiles:

여러 개의 분석 파일을 지정하여 분석하면 여러 개의 파일은 하나의 통합된 파일과 같이 단어회군의 빈도수를 분석한다. 그러나 여러 개의 파일을 개별적으로 분석하고 싶다면 BatchFiles을 선택하여 한 번에 여러 개의 파일을 개별 분석할 수 있다. 이런 경우 저장 과정은 따로 필요하지 않으며 분석 파일들에 대한 결과는 “분석파일명_collo.txt”라는 이름으로 분석 파일이 있던 폴더에 각각 저장된다.

1.3.4. 분석 수행

분석 파일과 저장 파일 지정 후 Settings 영역의 [process Files] 버튼을 클릭하면 단어회군 분석을 수행하게 된다.

1.3.5. 분석 수행 결과

예) BNC Written Sampler 분석 결과

Number of lines: 139,337 > 코퍼스를 구성하는 문장 수(마침표를 기준으로 함)

Number of words: 1,012,732 > 코퍼스를 구성하는 단어수

Number of collocations: 17,015 > 분석된 총 단어회군 총수로 토큰 값과 일치

COLLOCATION

LIST	TOKENS/%	TYPES/%	FAMILIES/%
one	6488/38.13	1104/16.26	466/10.08
two	2049/12.04	720/10.61	402/8.69
three	1324/7.78	590/8.69	360/7.78
four	936/5.50	470/6.92	315/6.81
five	803/4.72	410/6.04	290/6.27
six	689/4.05	362/5.33	262/5.66
seven	629/3.70	372/5.48	275/5.95
eight	446/2.62	278/4.10	215/4.65
nine	477/2.80	285/4.20	221/4.78
ten	483/2.84	285/4.20	223/4.82
11	401/2.36	265/3.90	207/4.48
20	255/1.50	208/3.06	169/3.65
Total	17015/100	6788/100	4625/100

> / 뒤의 수치는 비율을 의미함

Types Found In Collo List One > 위의 통계 결과 아래에는 각 등급에 포함된 단어회군 개별 하위 유형의 사용범위(Range)와 빈도수(Freq)가 제시되며 F1, F2 ... 는 여러 개의 코퍼스를 분석했을 경우 각 코퍼스에서 나타나는 단어회군 유형의 빈도수를 의미함

TYPE	RANGE	FREQ	F1
a few days later	1	4	4
a few hours later	1	1	1
a few weeks ago	1	2	2
a few weeks later	1	3	3
a few years ago	1	4	4

Types Found In Collo List Two

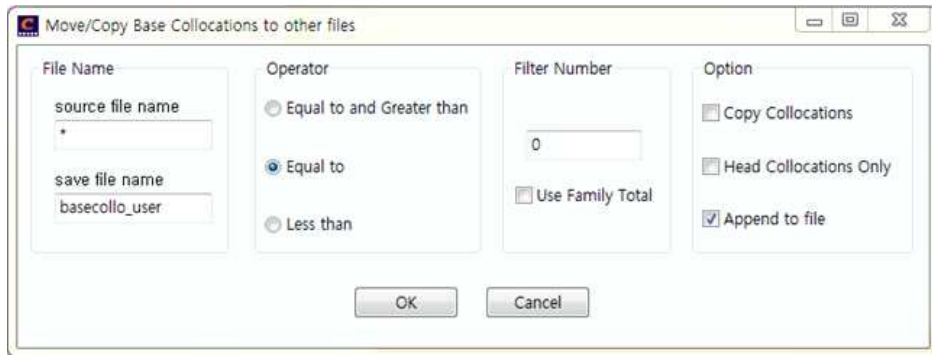
TYPE	RANGE	FREQ	F1
a couple of days ago	1	1	1
a couple of weeks	1	1	1
a few days ago	1	2	2
a few feet away	1	1	1
a long time ago	1	1	1
a lot of things	1	1	1
a wide range of	1	16	16

LIST OF FAMILY GROUPS > 다어휘군의 하위 유형(MWU type)의 등급별 사용범위와 빈도수 정보 다음에는 다어휘군(MWU family)의 사용범위와 빈도수 정보가 다시 제시되며 TYFREQ는 기본형으로 제시된 유형의 빈도를 의미하고 FAFREQ는 다어휘군 하위 유형들의 빈도수 총합을 의미하는 다어휘군의 빈도를 의미함

Families Found In Collo List ONE	RANGE	TYFREQ	FAFREQ	F1
a lot more	1	1	1	1
a lot of people	1	7	8	8
a lot of times	1	0	5	5
all of a sudden	1	6	8	8
all over	1	29	47	47
all right	1	43	43	43
also includes	1	6	10	10
around here	1	1	1	1
as well	1	214	214	214

1.3.6. Move/Copy Base collocations to other files

기존 Basecollo에 포함된 다어휘군 목록을 복사하여 새로운 다어휘군 목록을 개발할 때 사용하며 복사에 사용할 Basecollo의 수는 [그림 2]의 기본화면에서 Number of Basecollo Files에 입력한다.



[그림 2] 다어휘군 이동/복사 화면

Source file name:

다어휘군의 하위 유형을 Basecollo에서 복사하고자 한다면 아래 그림과 같이 기본형을 차례로 입력한 후 프로그램 폴더 안에 저장한다. 저장한 파일명을 source file name에 다시 입력한다.

Save file name:

다어휘군 하위 유형을 복사하여 저장할 목록의 이름을 지정한다. 기본값은 'Basecollo_user'로 설정되어 있다.

Source File	복사 후 저장한 결과 파일
at the last minute	5957 at the last minute 1475
all over	last minute 3806
	last minutes 431
	last few minutes 245
	32776 all over 19912
	all over again 2374
	all over the country 3196
	all over the world 4789
	all over the place 2505

Operator:

Equal to and Great than - Filter Number에 설정 빈도수를 입력하면 기존 Basecollo에 포함된 다어휘군 유형 중 그 빈도수와 같거나 큰 빈도수의 유형만을 복사한다.

Equal to - Filter Number에 설정 빈도수를 입력하면 기존 Basecollo에 포함된 다어휘군 유형 중 그 빈도수와 같은 유형만 복사한다.

Less than - Filter Number에 설정 빈도수를 입력하면 기존 Basecollo에 포함된 다어휘군 유형 중 그 빈도수 보다 작은 유형만 복사한다.

Filter Number:

Basecollo에 포함된 다어휘군 중 복사할 범위를 제한하고 싶을 때 설정하는 빈도수 입력란이다.

Use Family Total:

다어휘군의 개별 유형의 빈도수가 아닌 하위 유형의 총합 즉, 총 다어휘군 빈도수를 기준으로 복사의 범위를 제한할 때 사용한다.

Option:

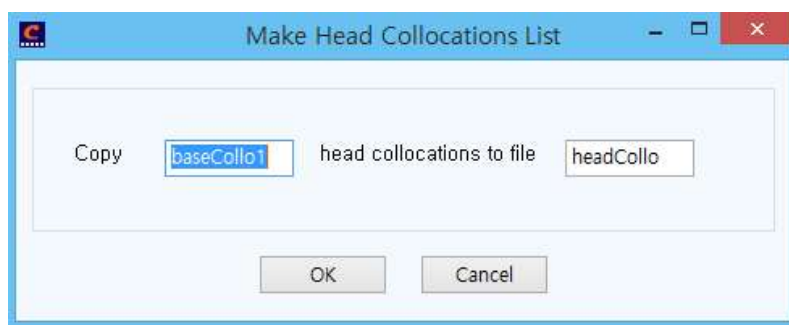
Copy Collocations - 조건을 만족하는 다어휘군 유형 또는 다어휘군을 기존 Basecollo 파일에서 “save file name”에 입력한 파일에 복사한다.

Head Collocations Only- 조건을 만족하는 기본형 즉 Head MWU 만을 “save file name”에 입력한 파일에 복사한다.

Append to file - 다어휘군 목록을 여러 번 복사할 때 선택하면 첫 번째 복사로 생성된 파일의 맨 아래 추가로 복사한 다어휘군 유형 또는 다어휘군이 연속하여 붙는다.

1.3.7. Make Head Collocations List

프로그램에 탑재된 특정 Basecollo 파일의 다어휘군 목록에서 기본형인 Head MWU만을 복사하여 새로운 파일에 저장한다.



[그림 3] 다어휘군 기본형 목록 생성 화면

입력란에 복사 대상의 Basecollo파일명을 지정하고 저장할 파일명을 입력하면 프로그램 폴더에 저장된다.

1.3.8. Remove duplicate collocations

Number of Basecollo Files에 입력한 다어휘군 목록을 대상으로 중복 다어휘군 유형 검사를 수행하여 원래 수정전 다어휘군 목록을 “Basecollos_org.txt”로, 중복되어 삭제된 다어휘군 항목만을 정리하여 “Basecollos_dup.txt” 로, 그리고 중복된 다

어휘군을 제거한 새로운 다어휘군 목록을 Basecollos_new.txt 으로 저장한다.
Basecollos_new.txt의 생성시 중복된 연다어휘군 유형은 앞선 유형을 남기고 뒤에
따르는 중복된 유형을 자동으로 삭제한다.

© **Citation:**

**Shin D., Chon, Y. V., Lee, S., & Park, M. (2018). *COCA_MWU20 ColloGram*
[Computer Software]. Seoul, South Korea: e-future.**