
ColloGram

User Guide (2018)

Developers

Shin, Dongkwang (Gwangju National University of Education)

Chon, Yuah. V. (Hanyang University)

Lee, Shinwoong (Hanyang University)

Park, Myongsu (Sangmyung University)

Table of Contents

1. Introduction	3
2. Program overview	3
2.1. Functions of ColloGram	4
2.2. Component Files of ColloGram	4
2.2.1 <i>File Construct of Collocation Lists</i>	4
2.3. Collocation Analysis Process	5
2.3.1 <i>File to Process</i>	6
2.3.2 <i>Save File</i>	6
2.3.3 <i>Analysis Options</i>	6
2.3.4 <i>Starting the Analysis</i>	7
2.3.5 <i>Analysis Results</i>	7
2.3.6 <i>Move/Copy Base collocations to other files</i>	9
2.3.7 <i>Make Head Collocations List</i>	10
2.3.8 <i>Remove duplicate collocations</i>	11

1. Introduction

This manual has been developed for user convenience and understanding of ColloGram, a program for analyzing the use of English collocations.

2. Program Overview

The analyses of MWUs in the existing MWU programs have often been based on the repetition of ‘N-gram’ patterns rather than a specific MWU list. In comparison, ColloGram named from the compound, Collocation and N-gram or Program, bases its analysis on a Multiword Unit (MWU) list from the Corpus of Contemporary American English (COCA) which so far (1990-2015) consists of a 5 hundred-million-word corpus. For the development of ColloGram, the corpus (4.5 hundred-million words) compiled during the period of 1990-2009, which became available in 2014 to the public, was utilized for the program. The MWU list contains 10,000 MWU families (31,680 MWU types) where the new notion of MWU Family is adopted. A MWU family includes its inflected forms (inflected verbs, singular/plural forms of nouns) and derived forms (different word order, forms with more or less constituents). The criteria to extract the COCA_MWU families were Min. Range 4 from the COCA 5 domains and Min. Frequency 20. The MWU family list topped on the ColloGram program is called **COCA_MWU20** where 20 indicates the number of graded 500-item bands. The functions of ColloGram are similar to those of RANGE, the vocabulary analysis program, by Heatley and Nation (2002). The program identifies MWUs where all the words are immediately adjacent to each other (that is, continuous MWUs, Max. 10 words). Finally, a subtractive method (Martinez & Schmitt, 2012) was adopted to produce a more accurate frequency figure of the MWUs in developing ColloGram. For example, **as opposed to** can be subsumed under **opposed to**, a head MWU. However, in order to obtain the exact frequency for **opposed to**, there was need to subtract the number of occurrences of the string **as opposed to** (1,615) from the number of times the bigram **opposed to** appears in the corpus (2,674). That is, the true frequency of **opposed to** is 1,059.

2.1. Functions of ColloGram

- Counts the frequency of collocation types, and families
- Copies Family Members (derivation and inflectional forms of the head) based on the head collocation
- Extracts head collocation list from the list
- Removes duplicate collocations in the list

2.2. Component files of ColloGram

This program provides collocation lists (Basecollo1.txt, Basecollo2.txt, Basecollo3.txt...Basecollo21) and an execution file (32bit, 64bit), available without an installation process.

- Basecollo1.txt~Basecollo16.txt: 1-15 include 500 collocation families, and 16 includes 130 collocation families
- Collogram32.exe: for general Desktop users
- Collogram64.exe: for the latest Laptop users

2.2.1. File Construct of Collocation Lists

The collocation lists consist of "Total frequency + Tab + head collocation (the most frequent collocation type) + 1 space + frequency of types. However, the program can be run without frequency data. Inflected forms (inflected verbs, singular/plural forms of nouns) and derived forms (different word order, forms with more or less constituents) are listed below the head collocation. Collocations in the Basecollo, including head collocations and all types, satisfy the minimum frequency of 20 in COCA, and also reach the minimum range of 2 in the 8 corpora. Subtypes consist of "Tab + collocation type + 1 space + frequency of the type". Thus when needing to add extra collocation types, the items need to be typed in by following the specific format. As seen below, the program can be run either by including frequency data or without it.

Including frequency data	Without frequency data
119854 as well 119854	as well
96337 years old 23243	years old
year old 73094	year old
76631 years ago 65359	years ago
year ago 7092	year ago
a few years ago 4180	a few years ago
65116 all right 65116	all right
57613 right now 57613	right now
57015 so much 48256	so much
not so much 4795	not so much
so much that 2086	so much that
so much to 1878	so much to

2.3. Collocation Analysis Process

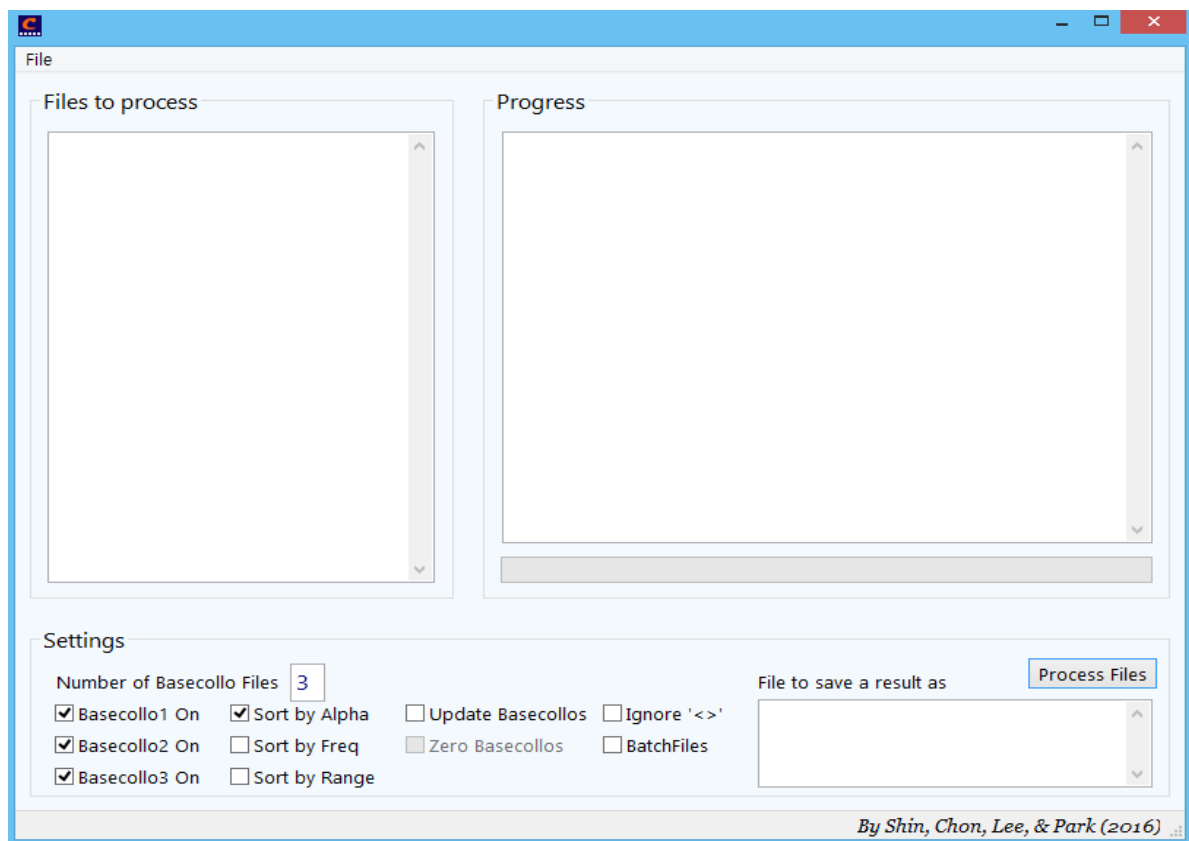


Figure 1 Screenshot of ColloGram

2.3.1. File to process

Choose the file to analyze from the [File]-[Open] Menu. More than one file can be chosen at once, or the files can also be dragged and dropped in the box.

2.3.2. Save file

Designate where you would want the analysis results to be saved at [File]-[Save] Menu. If not, click [Settings]-[BatchFiles]-[Process Files], and results will be saved as "originalfilename_collo.txt" in the same folder where the target file exists.

2.3.3. Analysis Options

Number of Basecollo Files:

Choose the number of collocation lists to be applied for analysis. For example, the number of Basecollos is set at three in Figure 1, but different numbers of Basecollos can be entered at the box of "Number of Basecollo Files" depending on the number of embedded collocation lists in the program. The program currently offers up to 16 collocation lists (Basecollo1~Basecollo16). More number of new Basecollo files can be applied in the analysis by adding more collocation files into the same folder where other Basecollo lists already exist.

Sort by Alpha:

This alphabetizes the results of collocation analysis.

Sort by Freq:

This sorts the results in the order of frequency.

Sort by Range:

This sorts the results in the order of range (measurement of how broadly the collocations have appeared in the text)

Update Basecollos:

Basecollos include information on total frequency of all head collocations and their sub-types based on the 8 corpora. By clicking [Update Basecollos], the frequency of collocations from the new corpus is added up with those in the existing list to be saved.

Zero Basecollos:

If you click [Zero Basecollos] when updating Basecollos, the existing frequency data is removed and replaced with the new frequency data. When the new file to process does not include previous collocation types, the frequency is shown as 0 (Zero).

Ignore '<>':

This excludes any data indicated as "< >". If the data contains any Tags, such as for indications on Parts of Speech, use of this function will allow only the original text to be analyzed.

BatchFiles:

When more than one file is chosen, the files are analyzed as a combined data set. To process several files separately, click [BatchFiles]. Files will be saved as "originalfilename_collo.txt" in the same folder where the target file exists.

2.3.4. Starting the Analysis

After selecting the file(s) to be analyzed and saved, click [Process Files] from [Settings] to start the analysis.

2.3.5. Analysis of Result

ex) Analysis result of BNC Written Sampler

Number of lines: 139,337 > **number of sentences in the corpus (based on periods)**

Number of words: 1,012,732 > **number of words in the corpus**

Number of collocations: 17,015 > **number of total collocations (number of tokens)**

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
16	34/ 0.20	27/ 0.40	27/ 0.58
Total	17015	6788	4625

Numbers following “/ ” refer to percentages

Types Found in Collo List One > **The statistics below show the collocation types included in the first collocation band with the range and frequency of each collocation type. F1, F2 ... refer to the frequency of the collocation type for each corpus (when analyzing one or more corpora).**

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

LIST OF FAMILY GROUPS > **After the range and frequency of ‘collocation types’ are presented, the range and frequency information are presented for the ‘collocation families’ at each collocation band. TYFREQ refers to the frequency of the head type alone and FAFREQ refers to the overall frequency of each collocation family (the head type plus family members). As such, TYFREQ can never be larger than FAFREQ.**

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

2.3.6. Move/Copy Base collocations to other files

As seen in Figure 2, this function is used to compile a new collocation list by copying the existing Basecollo collocations. Enter the number of Basecollos to be copied in [Number of Basecollo Files] as shown in Figure 1 above.

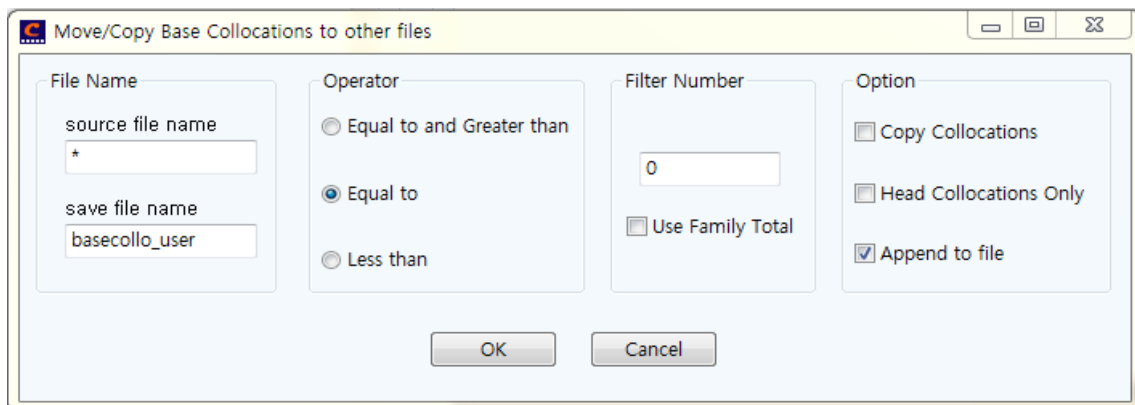


Figure 2 Screenshot of 'Move/Copy Base Collocations to other files'

Source file name:

To copy subtypes of collocations, enter head collocations in order as in the following table, and save to the program folder. Enter name for the saved file in [Source File Name].

Save file name:

Enter the name of the copied subtypes file. Default is 'Basecollo_user'.

Source File	Result file after copy and save
-------------	---------------------------------

years ago	76631 years ago 65359
all right	year ago 7092
so much	a few years ago 4180
	65116 all right 65116
	57015 so much 48256
	not so much 4795
	so much that 2086
	so much to 1878

Operator:

Equal to and Greater than- Filter Number: When a frequency cut off point is entered, collocation types of equal or greater frequency from Basecollo will be copied.

Equal to - Filter Number: When a frequency cut off point is entered, collocation types of equal frequency from Basecollo will be copied.

Less than - Filter Number: When a frequency cut off point is entered, collocation types of a lower frequency will be copied.

Filter Number:

Enter a frequency number to limit the range of collocations to be copied from Basecollo.

Use Family Total:

This function is used to restrict the range of collocations to be copied based on the frequency of families (total frequency of subtypes).

Options:

Copy Collocations – Copy the collocation types or families from Basecollo files that meet the conditions chosen in Figure 2 into the target file (the file typed in "save file name").

Head Collocations Only- Copy head collocations into the target file (the file typed in "save file name") only when they meet the conditions chosen in Figure 2.

Append to file – When copying collocation lists multiple number of times, newly copied collocations types or families will be listed below the collocations of the previous basecollo file.

2.3.7. Make Head Collocations List

Head collocations only can be copied from a selected Basecollo file in the program and saved as a new file (See Figure 3)

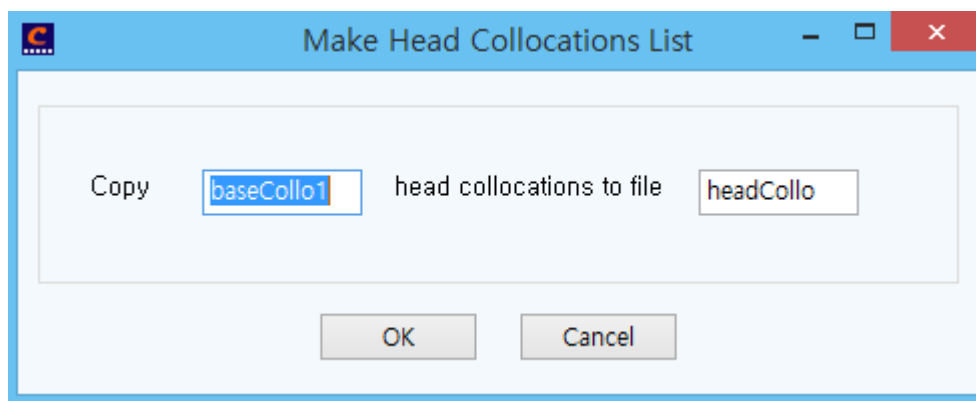


Figure 3 Screenshot of 'Make Head Collocation List'

Enter the name of Basecollo file to be copied, and the new file will be saved in the program folder.

2.3.8. Remove duplicate collocations

To remove duplicate collocations from the files in [Number of Basecollo Files], save the original collocation list (before removal of duplicate collocations) as "Basecollos_org.txt", the list of duplicate collocations as "Basecollos_dup.txt", and the new list without duplicate collocations as "Basecollos_new.txt". Among the duplicate collocations, the one that first appears is saved in the "Basecollos_dup.txt", automatically deleting the following duplicate collocations.

Citation:

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