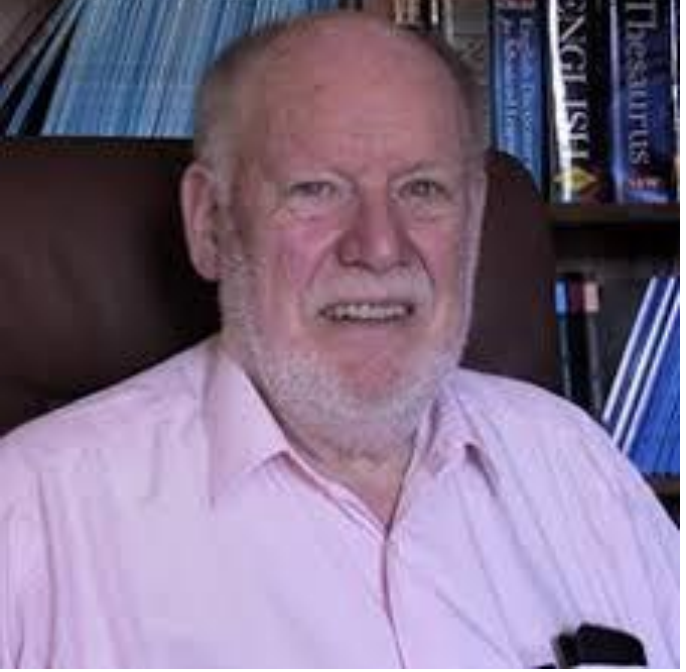
Collocation, part 2

- "Kun Marita Hakala kohtasi Loirin mustassa minihameessaan, se oli menoa."



Tyyppi A: John Sinclair + Birgmingham (UK)

Tyyppi B: Doug Biber + Flagstaff

LUMBERJACKS

- **Biber, Conrad, Reppen 1998: Corpus linguistics : investigating language structure and use: 2 korpuslingvistiikan päälähestymistapaa**

Table 1.1 *Association patterns in language use*

A. Investigating the use of a linguistic feature (lexical or grammatical)
   (i) Linguistic associations of the feature
      – lexical associations (associations with particular words)
      – grammatical associations (associations with particular grammatical constructions)
   (ii) Non-linguistic associations of the feature
      – distribution across registers
      – distribution across dialects
      – distribution across time periods

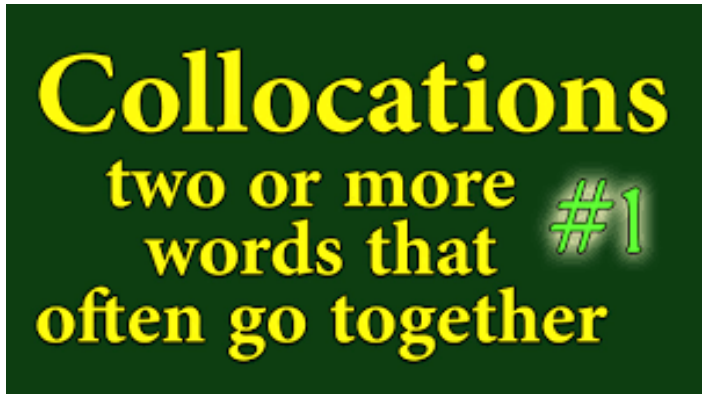
B. Investigating varieties or texts (e.g., registers, dialects, historical periods)
   (i) Linguistic association patterns
      – individual linguistic features or classes of features
      – co-occurrence patterns of linguistic features

Analysis steps

- Step 1: analyze the frequency of the pattern
- Step 2: analyze the contextual factors that influence its use

- …. And why was context important?

- Difference in context → difference in meaning
- Similar context → similar meaning
- Both for words, groups of words and constructions
- Many methods to quantify context
  - Collocation is one of them



- Tendency of two words to occur together
- Target word and its collocate occur together more frequently than what would be expected
- Collocates tell about

  →The target word meaning

  (Meaning is not *atomic* but defined by the target word + context)

  →Conventionalized ways of using language

2.6.2 *Immediate right collocates of* big, large, *and* great

Table 2.6 shows the most frequently recurring right collocates of *big*, *large*, and *great* in the academic prose and fiction samples from the Longman–Lancaster Corpus. A fuller analysis would require looking at the complete list of collocates; here we focus only on the top ten collocates in each register, excluding collocates that occur less than once per million words.

Academic prose (2.7 million words)

Big		Large		Great	
right collocate	freq. per million	right collocate	freq. per million	right collocate	freq. per million
enough	2.2	number	48.3	deal	44.6
traders	1.1	numbers	31.3	importance	12.5
		scale	29.4	number	8.9
		and	28.0	majority	8.1
		enough	15.9	variety	7.0
		proportion	11.8	extent	7.0
		amounts	10.7	part	4.1
		quantities	10.3	care	3.3
		part	10.0	advantage	2.6
		extent	8.9	detail	2.6
				interest	2.6

Fiction (3 million words)

Big		Large		Great	
right collocate	freq. per million	right collocate	freq. per million	right collocate	freq. per million
man	9.6	and	15.2	deal	40.4
enough	8.9	black	4.3	man	6.6
and	8.3	enough	3.6	burrow	5.6
black	8.3	house	3.0	big	4.6
house	7.6	room	2.7	aunt	4.3
one	7.0	white	2.7	care	4.0
toe	5.0	number	2.3	pleasure	4.0
old	4.6	for	2.3	and	3.0
red	4.3	man	2.0	relief	3.0
boy	3.6	one	2.0	black	2.7
room	3.6	in	2.0	to	2.7

- Michael Stubbs 1995: Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language*, 2, 1 (1995). Sections 1-3, 4.1, 5, 7
- Jantunen, J. H. 2009: Minulla on aivan paljon rahaa – Fraseologiset yksiköt suomen kielen opetuksessa [’I have really lots of money’ – Phraseological units in the teaching of Finnish]. *Virittäjä*, *113*(3). Noudettu osoitteesta https://journal.fi/virittaja/article/view/4202. Pages -368.

What is the objective of the study?

How are the terms collocation and semantic prosody defined? What about lemma?

What is the data used?

How is collocation analysis used?

Why are raw frequencies of collocates not enough?

Why is Stubbs worried about the representativeness of his dataset in Section 4.1?

What are the main results?

Semantic preference

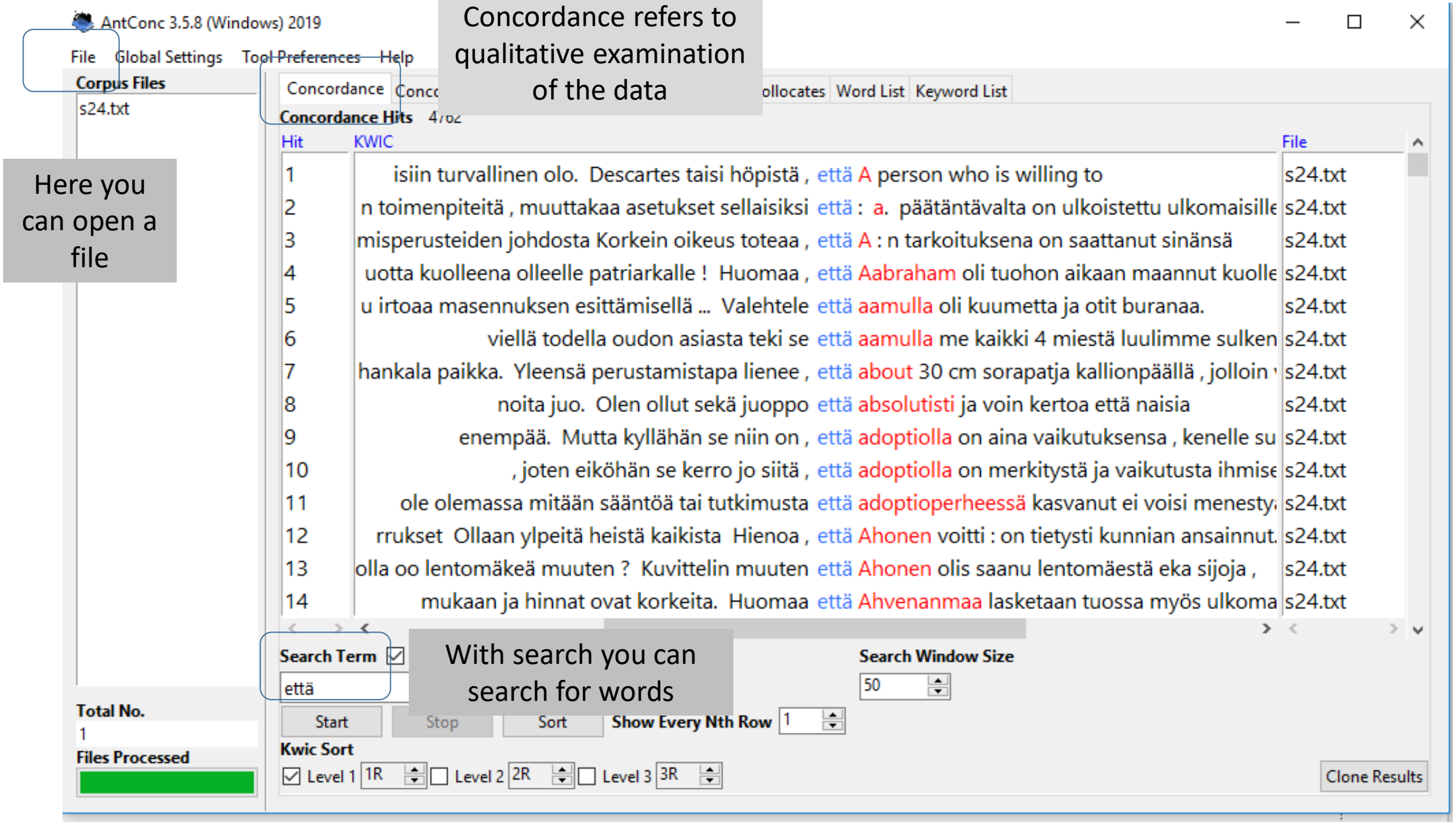
- Tendency of words to co-occur with semantically similar words
- Which semantic groups the collocates reflect?
- Stubbs (1995) reports that *cause* collocates with *abandonment* ‘luovuttaminen’, *accident* ‘onnettomuus’, *alarm* ‘huoli’ ja *anger* ‘viha’.
- → *cause* used in negative contexts

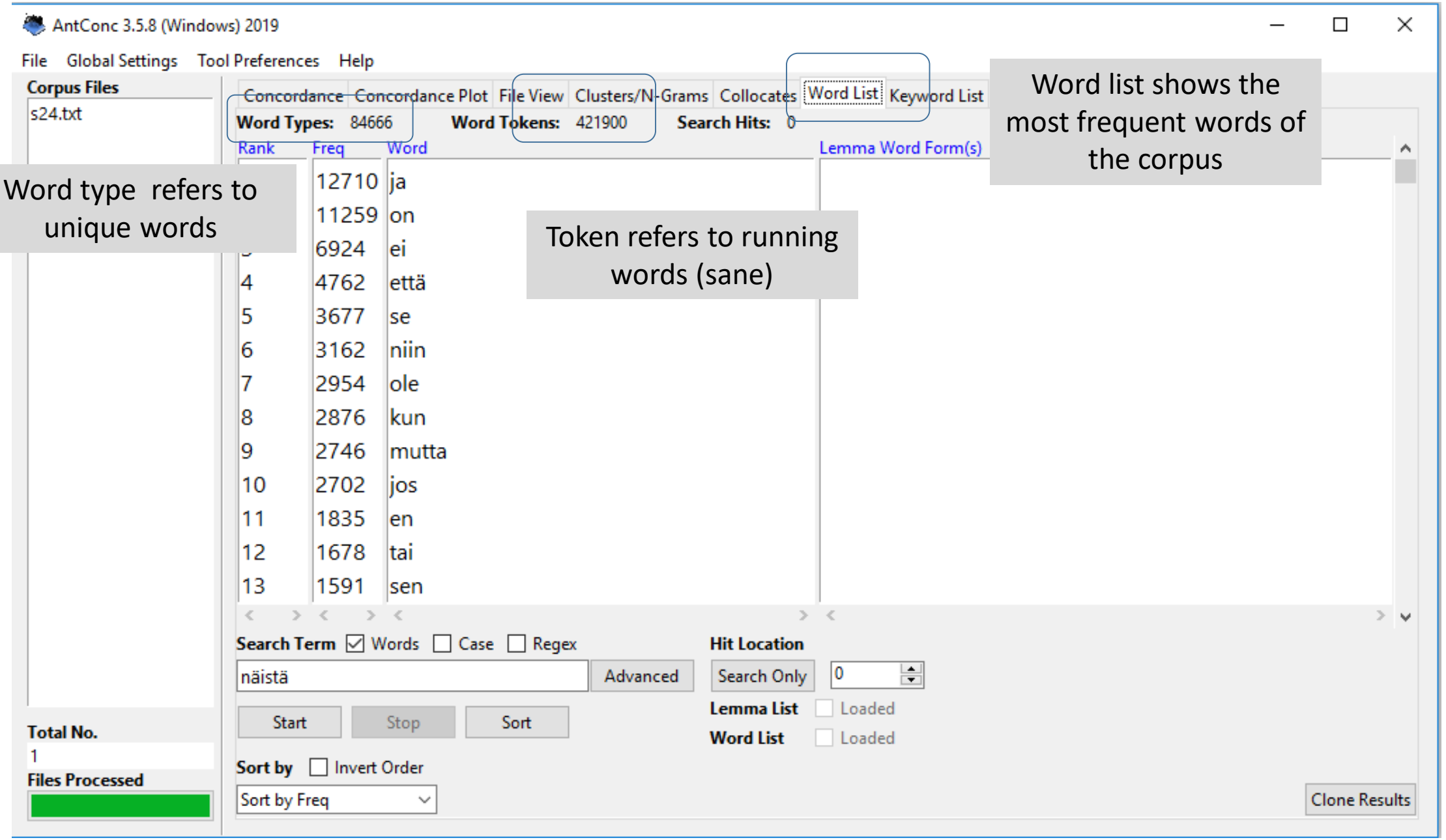
Semantic profile

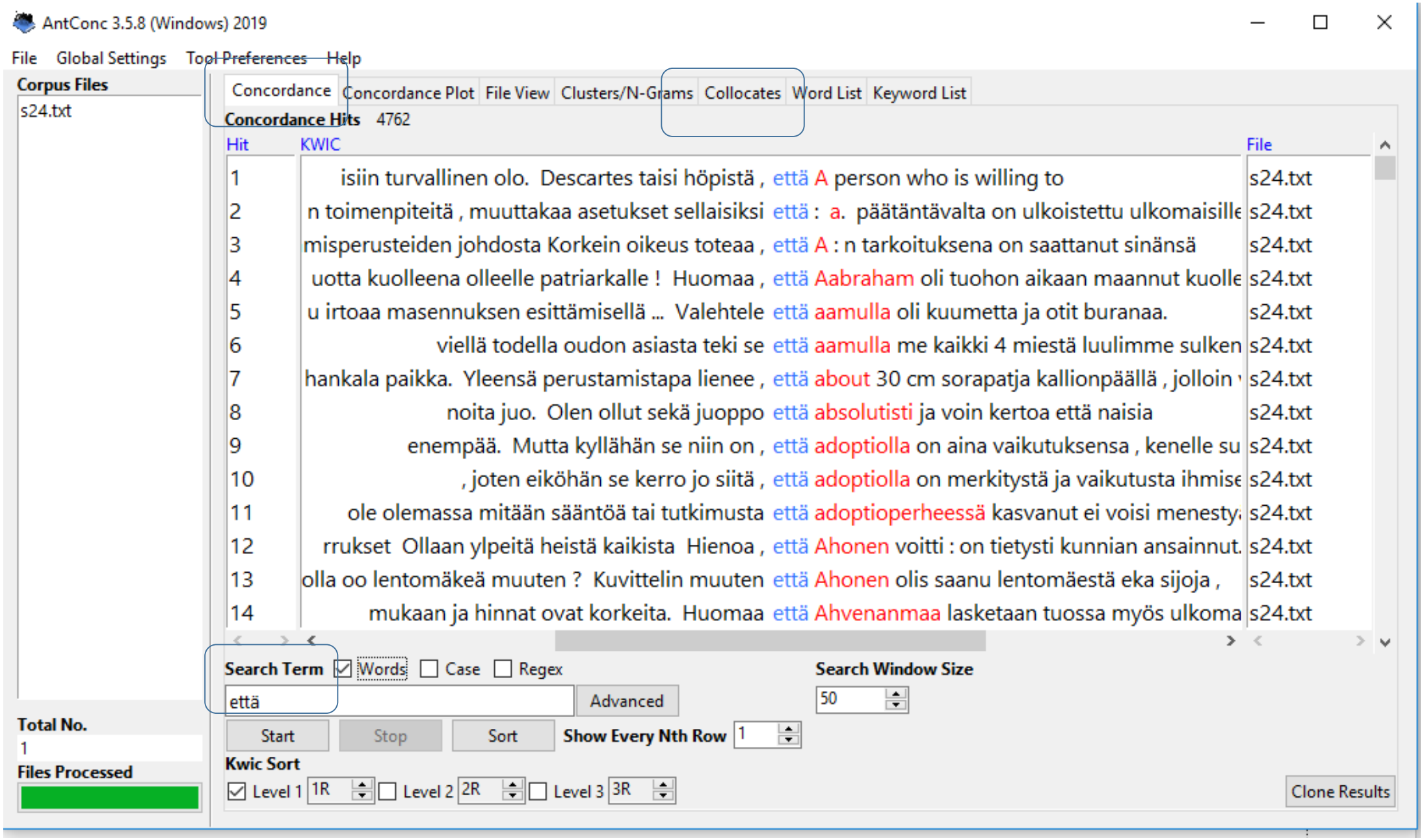
- Semantic groupings of the collocates
- Jantunen groups words collocating with *tässä* in research articles and in popular lifestyle texts
- Three groups of collocates
  - 1) Metatextual expressions: *tässä kirjassa, tässä luvussa, tässä tutkimuksessa*
  - 2) Point of view, perspective: *tässä perinteessä, tässä kohden*
  - 3) Abstract uses: *tässä tapauksessa, tässä vaiheessa*.
- Results tell about the use of *tässä* and its meaning more than what mere frequencies would
- Also register differences are clear
  - Metatextual expressions more frequent in academic texts → a typical way of structuring research articles

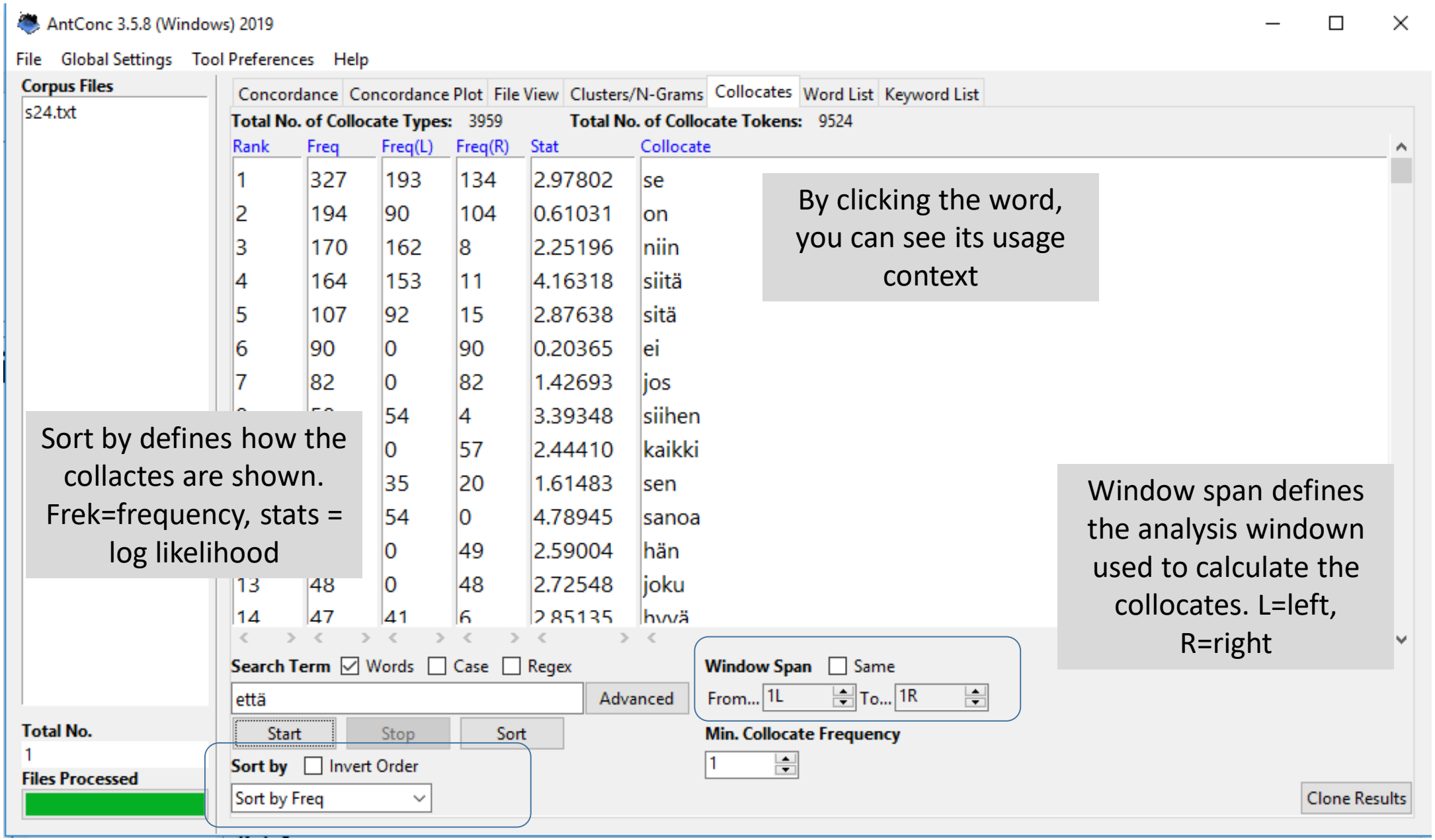
Let’s first look a video

- https://www.youtube.com/watch?v=9TsqFVrUYO0


AntConc 3.5.8 (Windows) 2019
File
Global Settings
Tool Preferences
Help
Concordance refers to qualitative examination of the data
Here you can open a file
Corpus Files
s24.txt
Concordance
Conc
ollocates
Word List
Keyword List
Concordance Hits 4702
Hit
KWIC
File
1 isiin turvallinen olo. Descartes taisi höpistä , että A person who is willing to s24.txt
2 n toimenpiteitä , muuttakaa asetukset sellaisiksi että : a. päätäntävalta on ulkoistettu ulkomaisille s24.txt
3 misperusteiden johdosta Korkein oikeus toteaa , että A : n tarkoituksena on saattanut sinänsä s24.txt
4 uotta kuolleena olleelle patriarkalle ! Huomaa , että Aabraham oli tuohon aikaan maannut kuolle s24.txt
5 u irtoaa masennuksen esittämisellä ... Valehtele että aamulla oli kuumetta ja otit buranaa. s24.txt
6 viellä todella oudon asiasta teki se että aamulla me kaikki 4 miestä luulimme sulken s24.txt
7 hankala paikka. Yleensä perustamistapa lienee , että about 30 cm sorapatja kallionpäällä , jolloin s24.txt
8 noita juo. Olen ollut sekä juoppo että absolutisti ja voin kertoa että naisia s24.txt
9 enempää. Mutta kyllähän se niin on , että adoptiolla on aina vaikutuksensa , kenelle su s24.txt
10 , joten eiköhän se kerro jo siitä , että adoptiolla on merkitystä ja vaikutusta ihmise s24.txt
11 ole olemassa mitään sääntöä tai tutkimusta että adoptioperheessä kasvanut ei voisi menesty s24.txt
12 rrukset Ollaan ylpeitä heistä kaikista Hienoa , että Ahonen voitti : on tietysti kunnian ansainnut. s24.txt
13 olla oo lentomäkeä muuten ? Kuvittelin muuten että Ahonen olis saanu lentomäestä eka sijoja , s24.txt
14 mukaan ja hinnat ovat korkeita. Huomaa että Ahvenanmaa lasketaan tuossa myös ulkoma s24.txt
With search you can search for words
Search Term
että
Search Window Size
50
Start
Stop
Sort
Show Every Nth Row
1
Kwic Sort
Level 1
1R
Level 2
2R
Level 3
3R
Clone Results
Total No.
1
Files Processed


AntConc 3.5.8 (Windows) 2019
File
Global Settings
Tool Preferences
Help
Corpus Files
s24.txt
Concordance
Concordance Plot
File View
Clusters/N-Grams
Collocates
Word List
Keyword List
Word Types: 84666
Word Tokens: 421900
Search Hits: 0
Rank
Freq
Word
Lemma Word Form(s)
12710 ja
11259 on
6924 ei
4 4762 että
5 3677 se
6 3162 niin
7 2954 ole
8 2876 kun
9 2746 mutta
10 2702 jos
11 1835 en
12 1678 tai
13 1591 sen
Search Term
Words
Case
Regex
näistä
Advanced
Hit Location
Search Only
0
Lemma List
Loaded
Word List
Loaded
Start
Stop
Sort
Sort by
Invert Order
Sort by Freq
Clone Results
Total No.
1
Files Processed
Word type refers to unique words
Token refers to running words (sane)
Word list shows the most frequent words of the corpus


AntConc 3.5.8 (Windows) 2019
File
Global Settings
Tool Preferences
Help
Corpus Files
s24.txt
Concordance
Concordance Plot
File View
Clusters/N-Grams
Collocates
Word List
Keyword List
Concordance Hits 4762
Hit
KWIC
File
1 isiin turvallinen olo. Descartes taisi höpistä , että A person who is willing to s24.txt
2 n toimenpiteitä , muuttakaa asetukset sellaisiksi että : a. päätäntävalta on ulkoistettu ulkomaisille s24.txt
3 misperusteiden johdosta Korkein oikeus toteaa , että A : n tarkoituksena on saattanut sinänsä s24.txt
4 uotta kuolleena olleelle patriarkalle ! Huomaa , että Aabraham oli tuohon aikaan maannut kuolle s24.txt
5 u irtoaa masennuksen esittämisellä ... Valehtele että aamulla oli kuumetta ja otit buranaa. s24.txt
6 viellä todella oudon asiasta teki se että aamulla me kaikki 4 miestä luulimme sulken s24.txt
7 hankala paikka. Yleensä perustamistapa lienee , että about 30 cm sorapatja kallionpäällä , jolloin s24.txt
8 noita juo. Olen ollut sekä juoppo että absolutisti ja voin kertoa että naisia s24.txt
9 enempää. Mutta kyllähän se niin on , että adoptiolla on aina vaikutuksensa , kenelle su s24.txt
10 , joten eiköhän se kerro jo siitä , että adoptiolla on merkitystä ja vaikutusta ihmise s24.txt
11 ole olemassa mitään sääntöä tai tutkimusta että adoptioperheessä kasvanut ei voisi menesty s24.txt
12 rrukset Ollaan ylpeitä heistä kaikista Hienoa , että Ahonen voitti : on tietysti kunnian ansainnut. s24.txt
13 olla oo lentomäkeä muuten ? Kuvittelin muuten että Ahonen olis saanu lentomäestä eka sijoja , s24.txt
14 mukaan ja hinnat ovat korkeita. Huomaa että Ahvenanmaa lasketaan tuossa myös ulkoma s24.txt
Search Term
Words
Case
Regex
että
Advanced
Search Window Size
50
Start
Stop
Sort
Show Every Nth Row
1
Kwic Sort
Level 1
1R
Level 2
2R
Level 3
3R
Clone Results
Total No.
1
Files Processed


AntConc 3.5.8 (Windows) 2019
File
Global Settings
Tool Preferences
Help
Corpus Files
s24.txt
Concordance
Concordance Plot
File View
Clusters/N-Grams
Collocates
Word List
Keyword List
Total No. of Collocate Types: 3959
Total No. of Collocate Tokens: 9524
Rank Freq Freq(L) Freq(R) Stat Collocate
1 327 193 134 2.97802 se
2 194 90 104 0.61031 on
3 170 162 8 2.25196 niin
4 164 153 11 4.16318 siitä
5 107 92 15 2.87638 sitä
6 90 0 90 0.20365 ei
7 82 0 82 1.42693 jos
54 4 3.39348 siihen
0 57 2.44410 kaikki
35 20 1.61483 sen
54 0 4.78945 sanoa
0 49 2.59004 hän
13 48 0 48 2.72548 joku
14 47 41 6 2.85135 hyvä
By clicking the word, you can see its usage context
Sort by defines how the collactes are shown. Frek=frequency, stats = log likelihood
Window span defines the analysis windown used to calculate the collocates. L=left, R=right
Search Term
Words
Case
Regex
että
Advanced
Window Span
Same
From... 1L
To... 1R
Start
Stop
Sort
Min. Collocate Frequency
1
Sort by
Invert Order
Sort by Freq
Clone Results
Total No.
1
Files Processed

Antconc, part 1

1. Install Antconc from https://www.laurenceanthony.net/software/antconc/
2. Load in a corpus, the soap.txt
3. Look at the most frequent words of the corpus. What are the 10 most frequent ones?
4. How many tokens (sane) does the corpus include? What about word types?
5. Let's analyze the use of *big*
6. Analyze in concordance line. How many occurrences?
7. Then count collocates
8. Try first with a window of L2 R2 and mimimum frequency of 5. How do the collocates look like?
9. Compare collocates sorted with frequency and stats. How are the results different? Can you guess why? Which option do you prefer?
10. Decrease minimum collocate frequency to 2. What happens? Can you guess why?
11. Increase the minimum frequency back to 5.
12. Analyze right and left collocates separately. How do they differ? Can you form functional groupings of the collocates? For this, you might want to read some texts too…

Antconc, part 2

- If you have time, do the same experiments with news.txt and compare the results