



UNIVERSITY OF  
EASTERN FINLAND

# Doing Lexical and Stylistic research using corpora

Corpus Linguistics Association of Nigeria (CLAN) workshop

**Temitayo Olatoye**

**Nov 21, 2025**

[temitayo.olatoye@uef.fi](mailto:temitayo.olatoye@uef.fi)

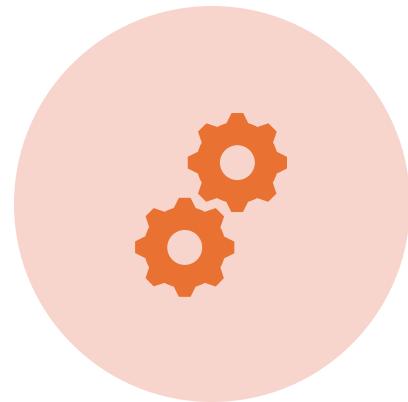
# Outline



TOOLS



USEFUL CONCEPTS



ADVANCED TEXT  
PROCESSING



# Tools

- ANTCONE: A corpus analysis toolkit for concordancing and text analysis

<https://www.laurenceanthony.net/software/antconc/>

- R and R Studio: Software and user-friendly integrated development environment (IDE)

<https://cran.r-project.org/>

<https://posit.co/download/rstudio-desktop/>

These tools are open-access.



# Concordancing in AntConc

A list of all the cases of a word, phrase or expression in a corpus

Target Corpus  
Name: ice\_nig\_spoken  
Files: 392  
Tokens: 622640

KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud ChatAI

Total Hits: 26 Page Size 100 hits 1 to 26 of 26 hits

	File	Left Context	Hit	Right Context
1	bnew_32.txt	centre Ibadan gave this advice during the funeral service of	Madam	Patience Aduke Ekundayo held at the church auditorium corresponden
2	bnew_32.txt	culture Abdulrauf Kamardeen NTA news the final funeral service for	Madam	Patience Aduke Ekundayo was witnessed by family friends and
3	bnew_32.txt	Reverend Suyi Falodun offered a prayer for the repose of	Madam	Patience Aduke Ekundayo stressing that the children of the
4	bnew_31.txt	Amuda Oju-Ere Ibadan by eleven a m on Saturday	Madam	Patience Ekundayo is survived by children grandchildren and great
5	bnew_31.txt	filed there this report the death has been announced of	Madam	Patience Ekundayo she passed away on first of March
6	bnew_08.txt	has the rest of the story the remains of late	Madam	Esther Arawuni Adewusi was laid to rest during the
7	bnew_08.txt	the brim today as friends relatives and children of late	Madam	Esther Arawuni Eluwara Adewusi gathered together to give her
8	leg_04.txt	fundamental is the issue of erm the harmonisation bill erm	Madam	should be in a position to brief us but
9	bnew_32.txt	Patience Aduke Ekundayo stressing that the children of the late	Madam	should see God as all sufficient in all solution
10	bnew_32.txt	paye Ibadan where Reverend Falodun encouraged the children of late	Madam	Aduke Ekundayo to follow her steps
11	cr_03.txt	this court what you know about this matter erm the	Madam	ANON erm came to the palace to report a
12	con_13.txt	they urinated imagine ah who will calculate dirty very dirty	Madam	ANON's class is for twelve twelve yes someone
13	con_02.txt	check up or so somebody said you have B P	Madam	B P means what so my son caused it
14	bnew_35.txt	right track participants at the colloquium include president of Liberia	Madam	Ellen Johnson-Sirleaf former executive secretary ECOWAS Mohammed
15	leg_10.txt	I believe in a cause and then just like erm	madam	has said I think we move some steps forward
16	btr_09.txt	I I would not feel happy so that's why	Madam	I probably you heard my shrill voice or my
17	con_55.txt	to have a feel that you are around I say	Madam	I'm always by there throughout ahem so at
18	unsp_01.txt	one hello my villagers can I plead on behalf of	madam	information on her comment towards we diasporans one thing
19	leg_11.txt	I don't know why because I know that erm	madam	is in a very good position to discuss our
20	bint_01.txt	Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq	madam	Time taken (creating KWIC results): 0.0198 sec

Search Query  Words  Case  Regex Results Set All hits Context Size 10 token(s)

madam Start Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Progress

# Text processing in R



# Natural Language Processing

## ▪ 1. Tokenization

This is the process of breaking text into smaller pieces called *tokens*.  
Tokens can be words, sentences, or even characters.  
Example: "I like oranges" → "I", "like", "oranges".

## ▪ 2. Lemmatization

This means converting a word to its *dictionary form* (called a lemma).  
It uses context and grammar. Thus, it produces real words.  
Example: "running", "ran" → "run".

## ▪ 3. Stemming

This is a simpler, rougher way of chopping words down to their *root form*.  
The result is not always a real word.  
Example: "running", "runner" → "run" or "runn".



## ■ 4. Stop-word removal

This is the process of removing words that do not add much meaning.

Examples: "the", "is", "and", "a".

## ■ 5. Part of Speech (POS) tagging

This means labelling each token with its grammatical role.

Examples: noun, verb, adjective, adverb.

*Cats sleep often.* → **Cats** (noun), **sleep** (verb), **often** (adverb), **.** (punctuation)



# Regular Expressions (REGEX)

**Regular expressions** (often called *regex*) are special patterns used to search for, match, or replace text.

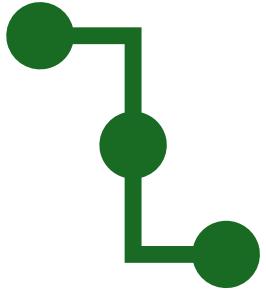
- They work like advanced search tools that let you describe what you want to find using symbols.

## Examples:

- The pattern [A-Z] matches any capital letter.
- The pattern cat|dog matches either “cat” or “dog”.
- The pattern \w+ matches a whole word.
- "(\\w+)/VERB(?:\\s+\\w+\\/\\w+){0,2}\\s+(\\w+)/ADP"  
#any verb with up to 2 intervening elements followed by a preposition

Regular expressions are rules that help you find specific patterns in text, not just exact words.

# Turning texts into social networks



How are characters in a text  
connected?



How can we visualize a complex  
object by its core structure?

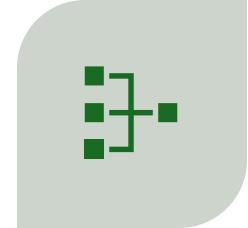


# Network Analysis

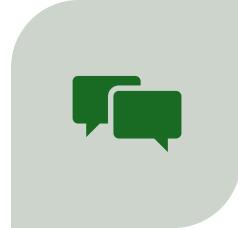
- Constructing and visualizing relationships in digital texts.
- Networks consist of nodes (dots) and edges (lines).

# Network Analysis

**Occurrences** – mentions of/references to a character

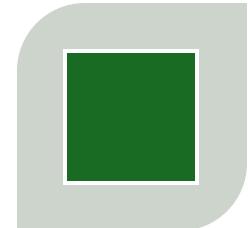


CO-OCCURRENCE  
NETWORK

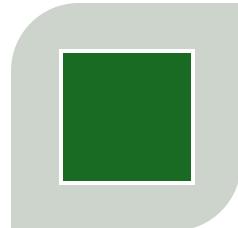


CONVERSATIONAL  
NETWORK

**Conversations** – character contributes/permforms an event



STATIC NETWORK

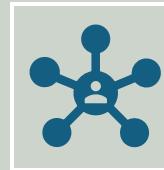


DYNAMIC  
NETWORK

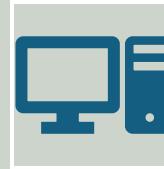
- Visualization of a whole - framing
- In sections – development over time

# Network Analysis

- Directed: the direction of the interaction matters.
- Undirected



Directed network



Undirected network



# Named Entity Recognition (NER)

- We tag the entities in a text.

**PERSON**, LOCATION etc.

- We group the co-referents in various parts of the text.
- We can also assign genders to the characters.

*Elizabeth, Lizzy, Miss Bennet* = Elizabeth Bennet

# Annotation

Mr. Bennet

She

Mrs. Long

His wife

Mr. Morris

~~His servants~~

Bingley

~~Our girls~~

he

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do not you want to know who has taken it?" cried his wife, impatiently.

"You want to tell me, and I have no objection to hearing it."

This was invitation enough.

"Why, my dear, you must know, Mrs. Long says that ~~Netherfield~~ is taken by a young man of large fortune from the north of England; that he came down on Monday in a chaise and four to see the place, and was so much delighted with it that he agreed with Mr. Morris immediately; that he is to take possession before Michaelmas, and some of his servants are to be in the house by the end of next week."

"What is his name?"

"Bingley."

"Is he married or single?"

"Oh, single, my dear, to be sure! A single man of large fortune; four or five thousand a year. What a fine thing for our girls!"



# References

- Anthony, Laurence. 2024. AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University.  
<https://www.laurenceanthony.net/software/AntConc>
- Ardanuy, Mariona Coll & Sporleder, Caroline. 2014. Structure-based clustering of novels. Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL 2014, pages 31–39.
- Bamman, David, Lewke, Olivia & Mansoor, Anya. 2020. An Annotated Dataset of Coreference in English literature. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 44 -54. <https://aclanthology.org/2020.lrec-1.6/>