First steps in FlexiConc: an initial case

In the hands-on session, we will use FlexiConc to study literary fiction.

This document serves as a step-by-step guide to the exercises.

Our case study deals with body part nouns in 19th century English novels and aims to find repeated patterns of language use. We use the 19C provided by the CLiC web app (Mahlberg et al. 2016).
The website can be accessed here:

https://clic-fiction.com/

This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

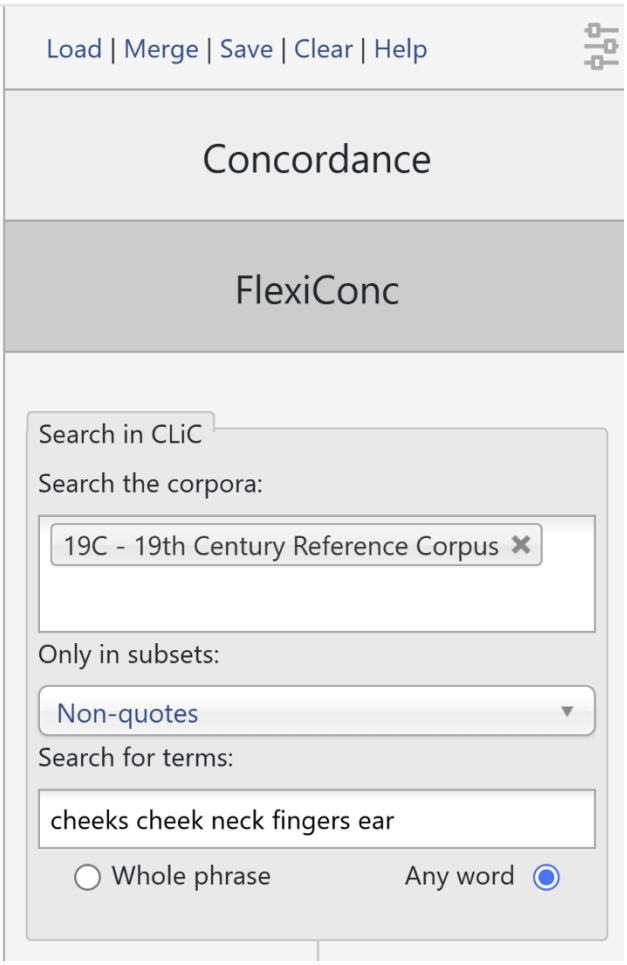
Part 1: technical steps

In this part, your task is to go through a pre-defined set analysis steps using FlexiConc in CLiC. Don't think too much about the concordance lines resulting from your steps just yet – you'll get a chance to do that in the second part.

Exercise 1: running a concordance search in FlexiConc

In this exercise, you will start your FlexiConc search. You will search for a range of body part nouns in *non-quotes*, that is, all parts of the 19C novels that are outside of quotation marks. All further exercises will be based on this initial concordance.

1. To use FlexiConc in CLiC, simply select the FlexiConc tab to the right of the page. The first step is to choose a corpus, a subset, and query terms.
2. Enter the search terms *cheeks, cheek, neck, fingers,* and *ear* **non-quotes** of the **19C** corpus.
3. Select the radio button *Any word* so that the words are searched separately rather than as a sequence.
4. After setting your search terms, click the **confirmation button** to start searching the corpus.
5. You will see the resulting concordance in the main window.



This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

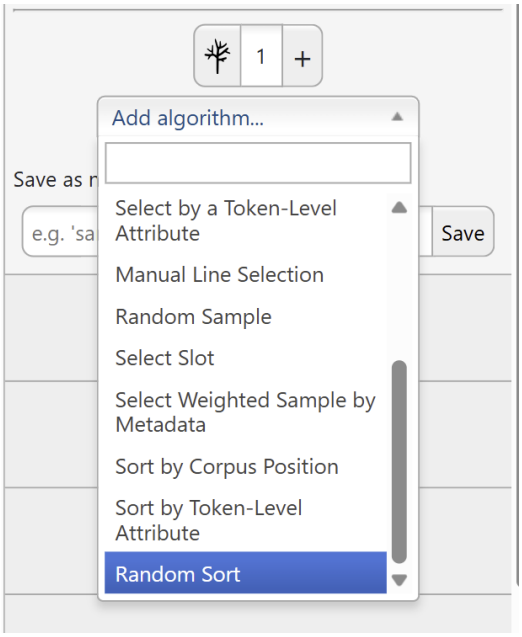
Exercise 2: a first overview – random sort

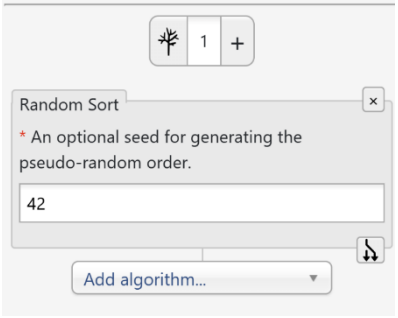
By default, the search terms are displayed in the following order:

I. The hits for each individual search term are shown first.
II. If there is more than one result, the hits are sorted alphabetically based on the name of the book they appear in.
III. If there is more than one result per search term and book, the hits are sorted based on their position within the book.

For our case study, we want a more general overview of how all of our search terms are used. For this, it can be useful to start with a random sort.

1. **Scroll down** in the FlexiConc panel until you see the **add algorithm** button below the tree symbol.
2. **Scroll down** in the selection menu to find the **random sort** function and click on it.
3. The algorithm will run immediately. Keep the **optional seed** that is set to 42 by default.





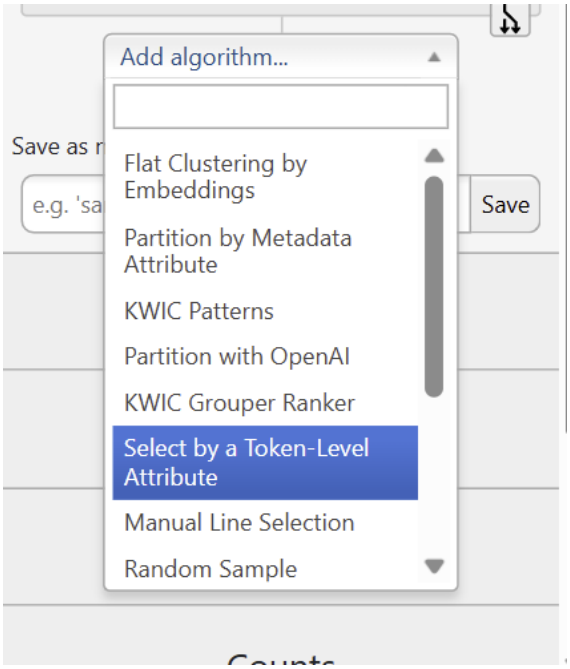
The **optional seed** number ensures that, every time someone runs the same 'random sort' step on the same data, they will be presented with the same ordering. This helps to make sure that results are **reproducible**.

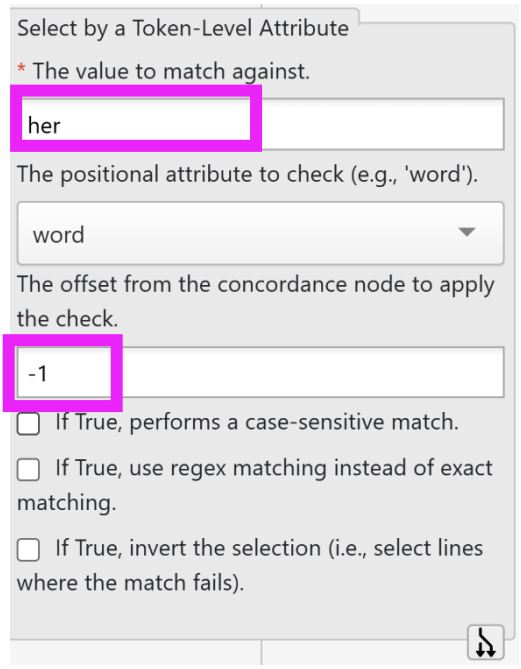
This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

Exercise 3: zooming in – select by a token-level attribute

Add an additional algorithm below the random sort by scrolling to ***select by a token-level attribute***. This step allows you to select only the concordance lines where the word in a specified position has a particular value. In this step, you will only focus on lines where the token immediately to the left of the node (= the body part noun) is *her*.

1. Set the **offset** to **-1**
2. Set the **value** to match against to **her**
3. All other settings remain unchanged





This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

Exercise 4: adding information – annotate with sentence transformers

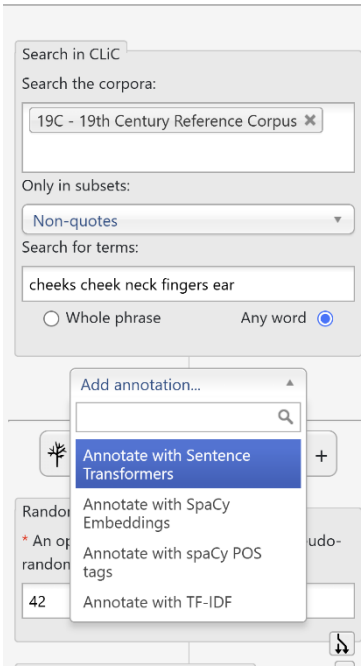
Sentence transformers (Reimers & Gurevych 2019) are a method to calculate the similarity between sentence pairs based on data from a large language model. In FlexiConc, we can use similarity scores based on these and other measures to determine the similarity between concordance lines.

In this step, we add the similarity information that is later used to perform clustering.

1. **Scroll up** to the query to find the **add annotation** menu.

❗ The annotation button is different from the *add algorithm* button – you won't see an immediate effect to your concordances in this step (except for some loading time).

2. Select **annotate with Sentence Transformers**
3. Click **confirm** in the concordance window.



This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

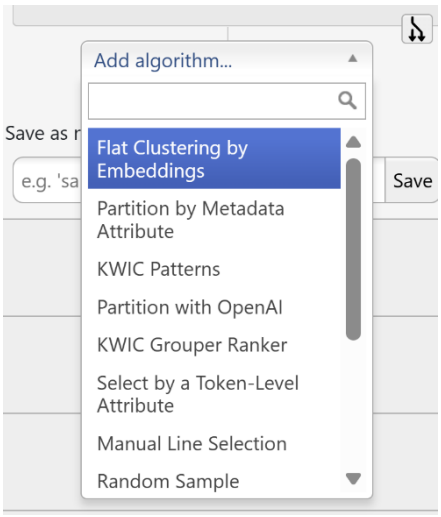
Exercise 5: Clustering – flat clustering by embeddings

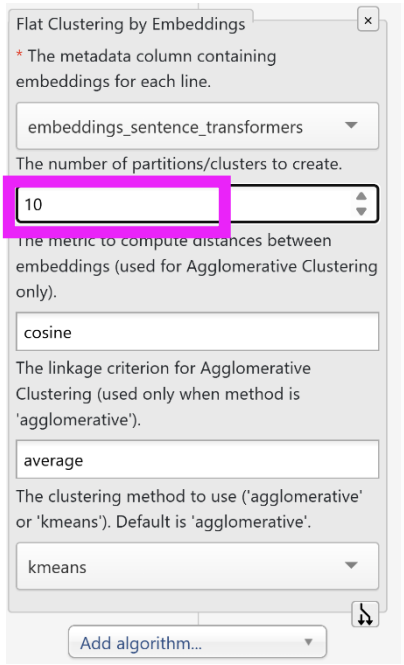
Clusters are groups of – in our case – concordance lines that are more similar to one another than to the rest of the lines in the concordance. In the methods that FlexiConc currently supports in CLiC, the analyst determines how many groups should be formed, and the algorithm produces exactly that number of line groupings – in our case, using the sentence embedding similarities from above.

Thus, the lines within a cluster tend to have high similarity scores to each other.

To perform such a clustering, scroll down to ***Add algorithm***.

1. Choose ***Flat clustering by embeddings***
2. Change the **number of clusters** to **10** and leave everything else as-is.





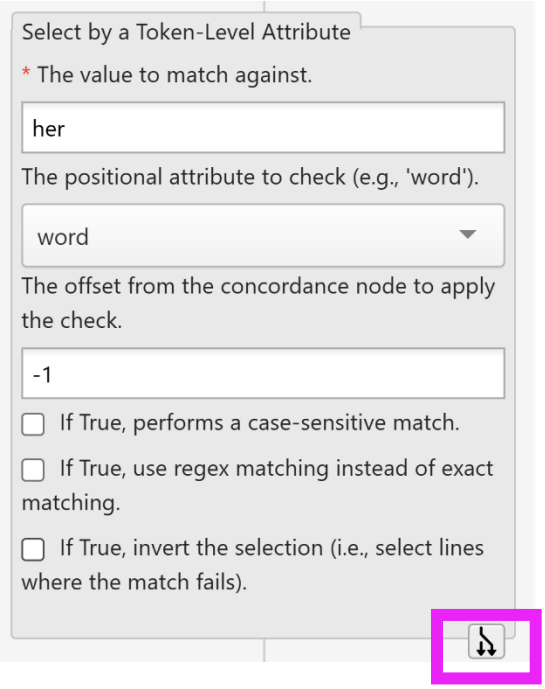
This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

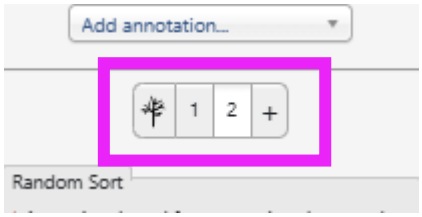
Exercise 6: branching out – moving up in the analysis tree

A key feature of FlexiConc is its analysis tree that documents all the algorithms we apply. This can be very useful during the analysis if we want to go back to a previous step and take alternative routes from there.

In this example, we will use this feature to examine *his* + *body-part noun* in the same way that we selected concordance lines for *her* + *body-part noun*.

1. **Scroll up** to the previous node *Select by a token-level attribute* where you selected *her*.
2. Click on the **branch** icon in the bottom right corner.



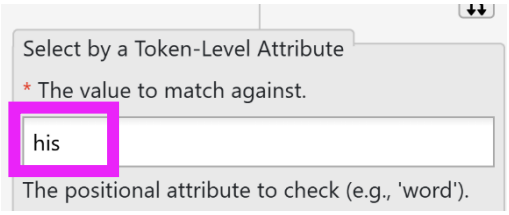


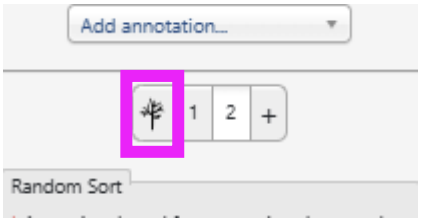
This step creates a **new branch** where all steps up until the one you branched off from are the same.

The following steps (in our case, *flat clustering by embeddings*) are not copied over from the 'old' branch. You can always go back to your previous branch by clicking **1** next to the tree symbol.

3. In the newly-created **branch 2** , **change the *Select by token value* to his**.
   On running this change, our branches will contain different selections of the overall concordance (*her cheeks / neck …* vs. *his cheeks/neck …*).
4. Click on the **tree symbol** to see a visualization of the entire analysis

This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

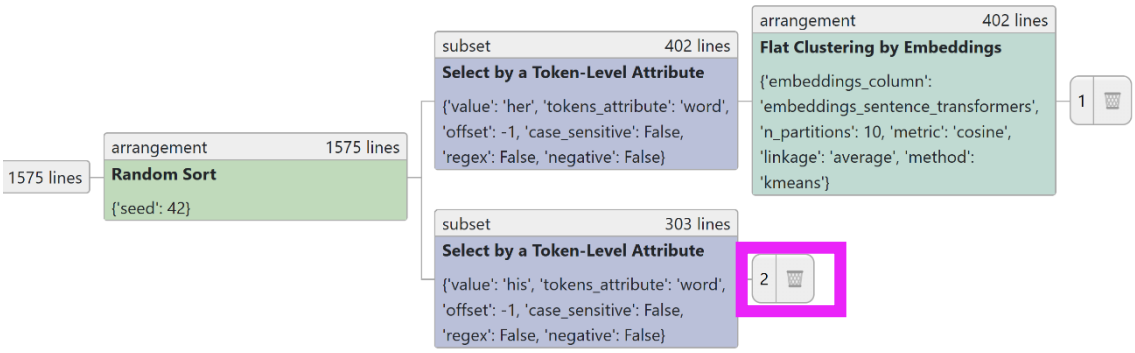




In the tree, you can click on any current branch to switch back to the analysis.

In our case, the *random sort* step is shared by both branches. Branch 1 additionally has the clustering step that we just applied.

5. click on **2** to go back to the branch that you just created (containing lines with *his* to the left of the node).



Exercise 7: another branch, another partition – clustering the new results

In this step, we will add an equivalent clustering step to our new branch 2. This will allow us to identify groups in the use of body part nouns for both of our selections.

1. Add the ***Flat clustering by embeddings*** ***algorithm*** *to branch 2.*
2. **Cluster the lines** using default settings (5 partitions).

Since the similarity scores based on sentence embeddings were calculated for the entire concordance, you **don't need to add a new annotation layer**! – this is why *annotation* is not the same as an *algorithm*.

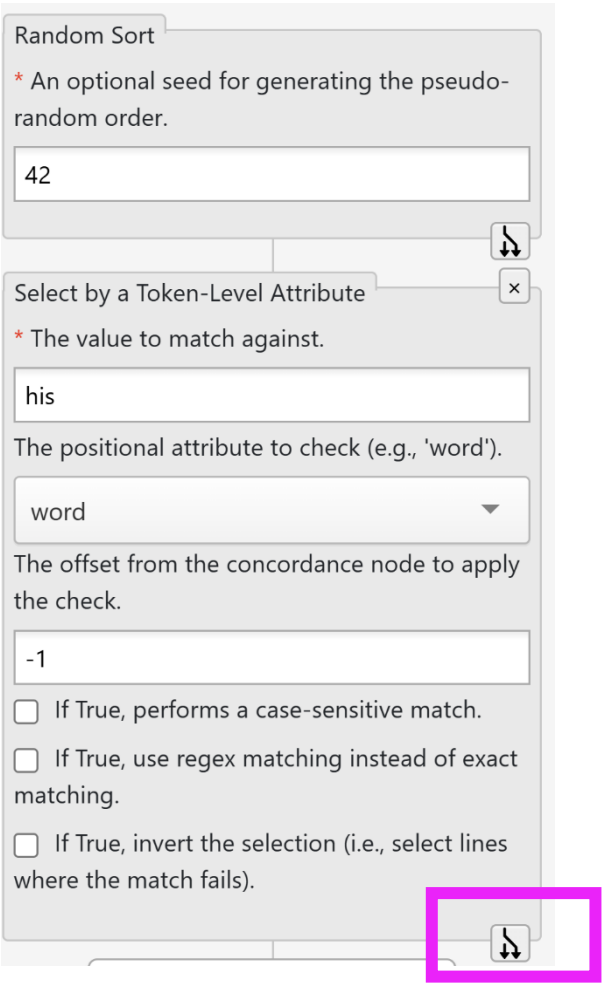
This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

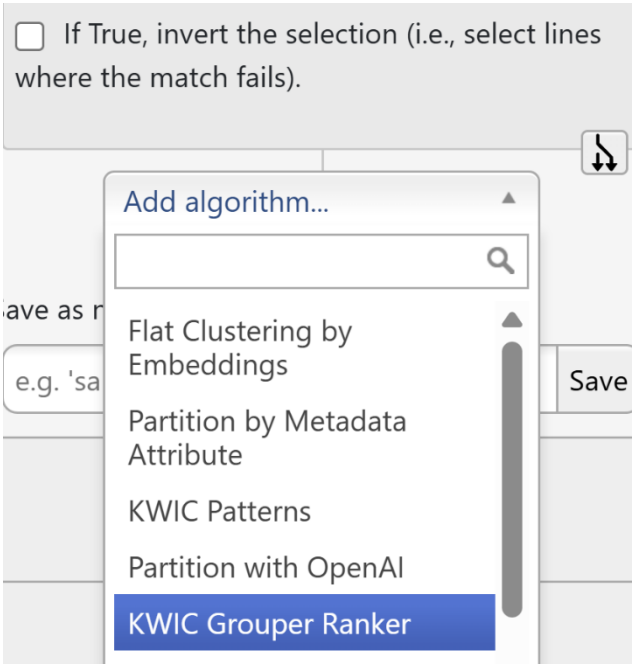
Exercise 8: good things come in threes – a new branch for *his* using KWICGrouper

KWICGrouper Ranker (a variant of the original KWICGrouper in CLiC, O'Donnell 2008) counts the occurrences of specified search terms and ranks the concordance lines based on the number of search matches.

In this example, we are interested in lines with repeated ***-ing* forms**.

1. Create a **new branch** under the selection for *his*
2. Select **KWICGrouper Ranker**





This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

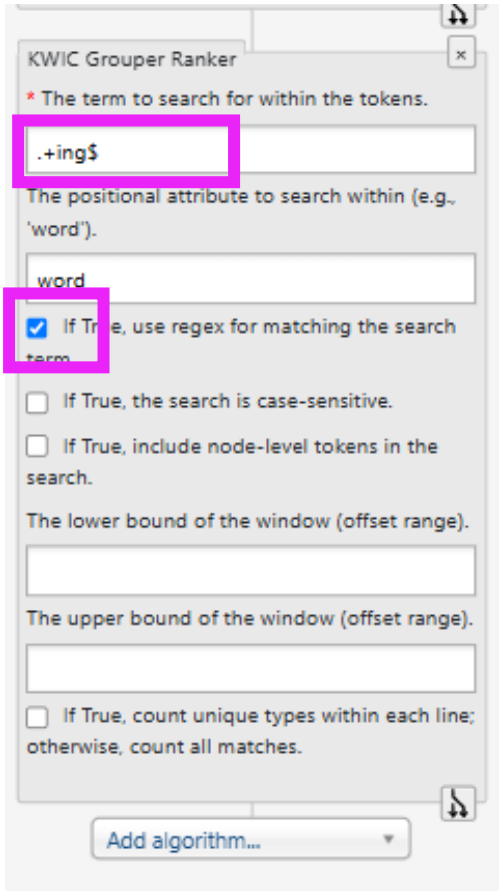
3. **Enable the checkbox for regex search**

4. In the search window, type **.+ing$**

This is a **regular expression** (*regex*). Regex allow you to specify patterns without spelling out the exact words that we want to match – since we're interested in *all* -ing forms, spelling them out explicitly could never give us a complete list!

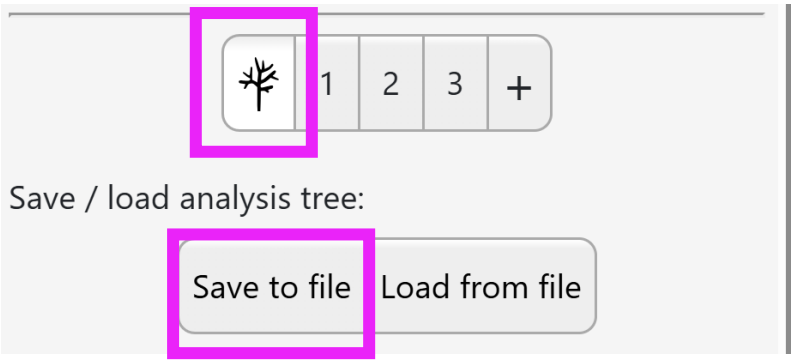
.+ means that the word can start with any character and be of arbitrary length; but there must be at least one character before *ing*.

$ means that the word has to end after *-ing*, so we don't get matches like *thingamy*. We will, however, still get some unintended results (e.g. *something*).



Exercise 9: last but not least

1. Go back to the **analysis tree** view
2. **Click on save to file**.
3. Name your tree **rc21_training_day_example.json**
4. Save it in a location where you'll be able to find it again!



This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

Bonus

[JupyterNotebook to do the same]

Part 2: interpretation

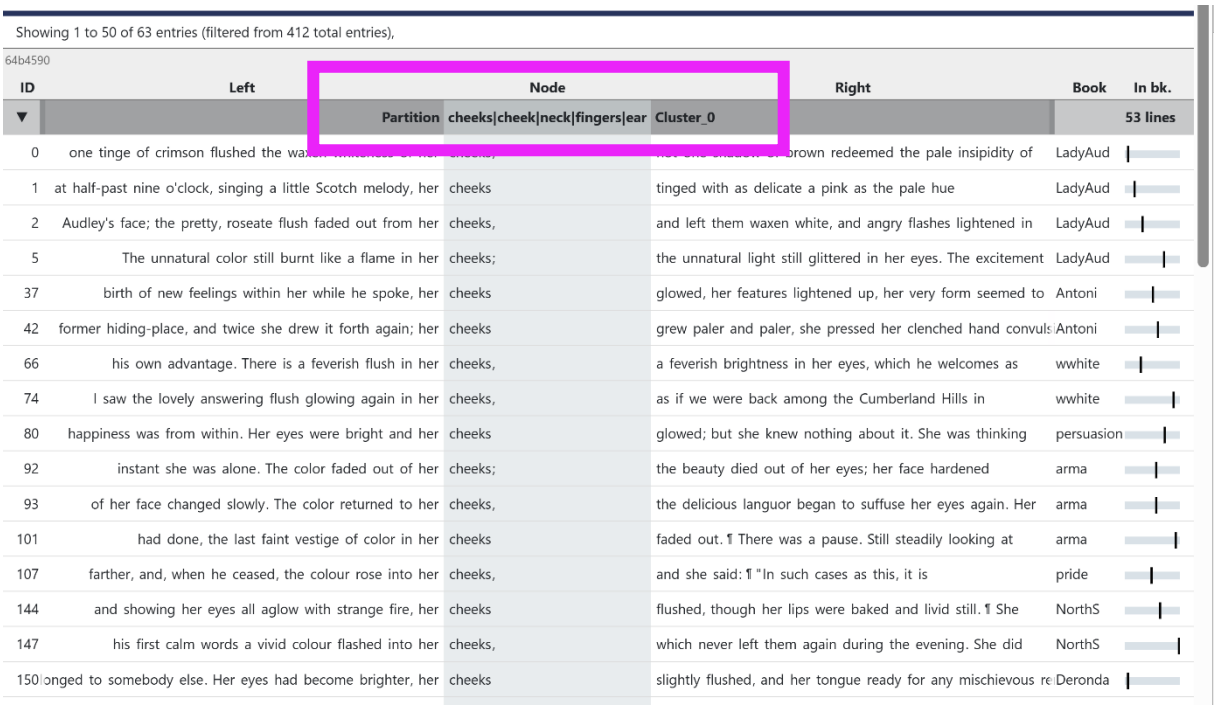
In this part, you'll look at selected concordance views from the steps that you just applied and try out some additional options. Don't worry if you don't make it through all steps – if you ***save your analysis tree to a file****, you can load it back up anytime to recover your steps!*

Exercise 1: clusters for *her* + body part noun

1. Go to **branch 1** of your analysis tree. To access it, you can load the JSON file for the tree that we just created to get there. If that doesn't work, use this link.

By default, the overview shows you the concordance lines in cluster_0, which is the first partition.

2. **Click on the partition label** for a given cluster to **collapse** and **un-collapse** the lines.



Showing 1 to 50 of 63 entries (filtered from 412 total entries),

64b4590

ID	Left	Node		Right	Book	In bk.
▼		Partition	cheeks|cheek|neck|fingers|ear	Cluster_0		53 lines
0	one tinge of crimson flushed the wa			rown redeemed the pale insipidity of	LadyAud	
1	at half-past nine o'clock, singing a little Scotch melody, her	cheeks		tinged with as delicate a pink as the pale hue	LadyAud	
2	Audley's face; the pretty, roseate flush faded out from her	cheeks,		and left them waxen white, and angry flashes lightened in	LadyAud	
5	The unnatural color still burnt like a flame in her	cheeks;		the unnatural light still glittered in her eyes. The excitement	LadyAud	
37	birth of new feelings within her while he spoke, her	cheeks		glowed, her features lightened up, her very form seemed to	Antoni	
42	former hiding-place, and twice she drew it forth again; her	cheeks		grew paler and paler, she pressed her clenched hand convuls	Antoni	
66	his own advantage. There is a feverish flush in her	cheeks,		a feverish brightness in her eyes, which he welcomes as	wwhite	
74	I saw the lovely answering flush glowing again in her	cheeks,		as if we were back among the Cumberland Hills in	wwhite	
80	happiness was from within. Her eyes were bright and her	cheeks		glowed; but she knew nothing about it. She was thinking	persuasion	
92	instant she was alone. The color faded out of her	cheeks;		the beauty died out of her eyes; her face hardened	arma	
93	of her face changed slowly. The color returned to her	cheeks,		the delicious languor began to suffuse her eyes again. Her	arma	
101	had done, the last faint vestige of color in her	cheeks		faded out. ¶ There was a pause. Still steadily looking at	arma	
107	farther, and, when he ceased, the colour rose into her	cheeks,		and she said: ¶ "In such cases as this, it is	pride	
144	and showing her eyes all aglow with strange fire, her	cheeks		flushed, though her lips were baked and livid still. ¶ She	NorthS	
147	his first calm words a vivid colour flashed into her	cheeks,		which never left them again during the evening. She did	NorthS	
150	onged to somebody else. Her eyes had become brighter, her	cheeks		slightly flushed, and her tongue ready for any mischievous re	Deronda	

This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

Left	Node	Right	Book	In bk.
Partition	cheeks|cheek|neck|fingers|ear	Cluster_0		53 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_1		31 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_2		60 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_3		43 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_4		27 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_5		45 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_6		34 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_7		35 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_8		53 lines
Partition	cheeks|cheek|neck|fingers|ear	Cluster_9		21 lines

Look at the concordances for some of the clusters

1. What are the differences between **clusters 0 and 4**?

2. How would you describe these similarities compared to patterns that you might identify through other means such sorting?

3. Change the number of clusters to a) 5 and b) to 15. What changes do you see? Which number of clusters seems to work best?

Exercise 2: clusters for *his* + body part noun

Go to **branch 2** of your analysis tree (link) to find the clusters for *his* + *body part noun*.

1. How do the uses of **cheek in cluster 0** compare to what you saw for *her cheek(s)*?
2. **Add a second annotation layer with spaCy embeddings** and
3. **Create a partitioning** that uses these new embeddings (default settings).

❗ **Don't add a new algorithm** to cluster anew – replace `embeddings_sentence_transformers` with `embeddings_spacy` in the active algorithm.

4. Do you see a difference between the two clustering algorithms? Which works better here?

This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).

Exercise 3: KWICGroups for *his/her* + body part noun

Go to **branch 3** of your analysis tree (link) to find the clusters for *his* + *body part noun*.

1. What kinds of features are highlighted through this use of KWIC Grouper? Do you see functional similarities between highly-ranked lines?
2. Apply the same KWICGrouping algorithm to the selection of her+body part noun – what do you see there?

Part 3: Recap questions

1. Which finding about body language surprised you most?
2. How might you continue this analysis?
3. What were your main learnings from this session?

Feedback: help us take FlexiConc further!

Please do give us feedback on our tool! We'd really appreciate if you could participate in our feedback survey here:

https://forms.gle/MZrqTPLgrPwbiFG89

To cite our work:

Dykes, Nathan; Evert, Stephanie; Mahlberg, Michaela; Piperski, Alexander (2025) First steps in FlexiConc: an initial case [Blog post]. RC21 blogs. Friedrich-Alexander-Universität Erlangen-Nürnberg.

References

Mahlberg, M., Stockwell, P., Joode, J. D., Smith, C., & O'Donnell, M. B. (2016). CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. Corpora, 11(3), 433-463.

O'Donnell, M. B. (2008). KWICgrouper–Designing a tool for corpus-driven concordance analysis. *International Journal of English Studies*, *8*(1), 107-122.

Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).

This handout builds on work funded by the "Reading Concordances in the 21st Century" research project supported by the Arts and Humanities Research Council (AHRC) (grant references: AH/X002047/1 & AH/X002047/2) and the Deutsche Forschungsgemeinschaft (DFG) (grant reference: 508235423).