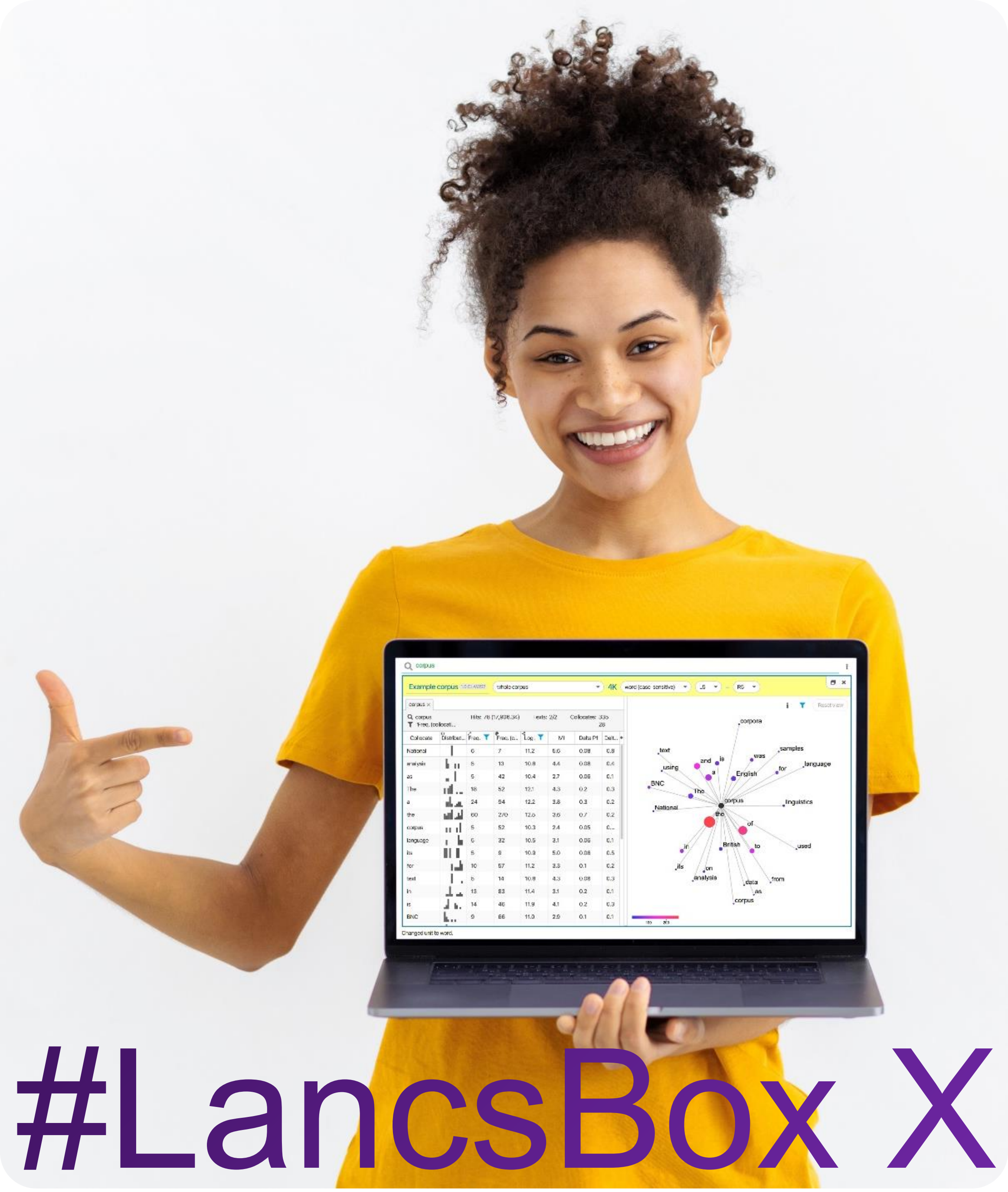


corpus Hits: 78 (17,808.34) Texts: 2/2 Collocates: 335

Freq. (collocat... 28

Collocate	Distribut...	Freq.	Freq. (c...	Log...	MI	Delta P1	Delt...
National		6	7	11.2	5.6	0.08	0.8
analysis		5	13	10.8	4.4	0.08	0.4
as		5	42	10.4	2.7	0.06	0.1
The		18	52	12.1	4.3	0.2	0.3
a		24	94	12.2	3.8	0.3	0.2
the		60	270	12.6	3.6	0.7	0.2
corpus		5	52	10.3	2.4	0.05	0...
language		6	32	10.5	3.1	0.06	0.1
its		5	8	10.8	5.0	0.08	0.5
for		10	57	11.2	3.3	0.1	0.2
text		5	14	10.8	4.3	0.08	0.3
In		13	83	11.4	3.1	0.2	0.1
is		14	46	11.9	4.1	0.2	0.3
BNC		9	66	11.0	2.9	0.1	0.1

Changed unit to word.

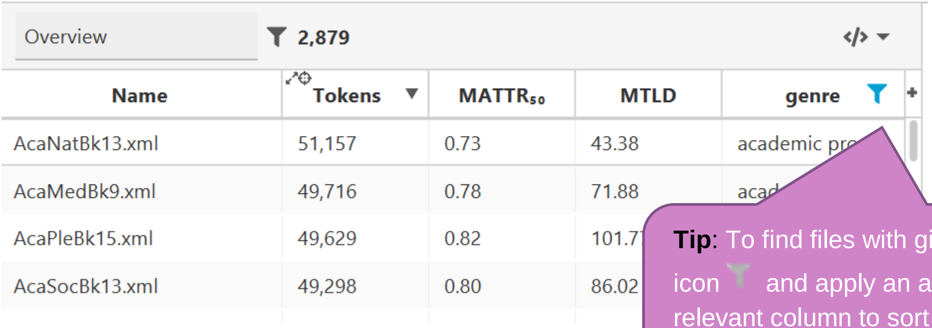
#LancsBox X

Text

Task 1 Corpus overview

In this task, you will explore the properties of individual texts in a corpus. Go to the Text tool in #LancsBox X and select the BNC2014 corpus (whole corpus). Provide the following information.

1. Number of files in the BNC2014 is ________88,171________.
2. The **largest** file has ________123,259________ tokens.
3. The **smallest** file has ________30________ tokens.
4. The number of files that are equal or larger than 10,000 words is ______1,820________.
5. The most lexically diverse file in **Academic prose** is _AcaMedRv88.xml_ with MATTR$_{50}$ _______0.88______.
6. The least lexically diverse file in **Informal speech** is _Sp0m2f99.xml_ with MATTR$_{50}$ _______0.64______.



Overview 2,879

Name	Tokens	MATTR₅₀	MTLD	genre
AcaNatBk13.xml	51,157	0.73	43.38	academic pr
AcaMedBk9.xml	49,716	0.78	71.88	acad
AcaPleBk15.xml	49,629	0.82	101.7	
AcaSocBk13.xml	49,298	0.80	86.02	

Tip: To find files with given properties click on the filter icon and apply an appropriate filter . Then click on a relevant column to sort files. Columns can be added by clicking on the + sign.

Lexical diversity

There are a number of lexical diversity measures showing the range of different words in a text. For the comparison of files of varying sizes, we need to go beyond a simple Type/token ratio (TTR) and compute more sophisticated measures such as Moving average type/token ration (MATTR) or a Measure of textual lexical diversity (MTLD).

Type/token ratio (TTR) expresses the proportion of types relative to the proportion of tokens. It is calculated by dividing the number of types in a text or corpus by the number of tokens. It decreases with text size so it cannot be used to compare texts of different sizes in a corpus.

Moving average type/token ration (MATTR) is calculated by dividing a text into standard sized overlapping segments (e.g. 50 words in MATTR$_{50}$) as a window moves through the file one token at a time. TTR is calculated for each overlapping segment and then the mean value of the TTRs is taken. MTTR is suitable for comparing texts of different sizes.

Measure of textual lexical diversity (MTLD) is the mean number of words in a text that maintain a given TTR value of .72.

Task 2 Distribution of linguistic features in texts

In the BNC2014, search for occurences of the past tense using the smart search `PAST_TENSE` (don't' forget to include the underscore). Answer the following questions:

1. In how many texts does the **past tense** occur? ____________81,397_______________
2. In how many texts does the **past tense** occur with a relative frequency that is higher than the average

relative frequency for the whole corpus?__Number of texts: _______________29,024 _________________

3. In how many newspaper texts does the past tense occur **at least once**? ___________46,112 _________
4. In how many newspaper texts does the past tense **not occur at all**? ______________4,098 _____________

Task 3 Analysing individual texts

In the BNC2014, find the text with the largest relative frequency of the search term `fuck*.` Provide information about this text.

1. Name of the text file: ______MagCla1338.xml_____
2. Genre: _______magazines______
3. Source: _______Classic Rock________
4. The swearword `fuck*` occurs ____13_______times in the text, which has ____148____tokens.

   This means on average, an f-word occurs every ___11.38____ words.

5. The function of the swearwords in this context is

   To describe songs on an album which contains a high frequency of the swearword fuck*'. The article reports the release of the album and also highlights the frequent use of the term in an individual song.

Developed at
Lancaster
University