Open research group

ESRC Centre for Corpus Approaches to Social Science

Lancaster University

- Open space for ideas
- Corpus linguistics and statistics
- Research community

Topics

Wednesday 16 October 12.00pm - 12.50pm UK time, **Statistics and language analysis - #LancsBox KWIC**

Wednesday 30 October 12.00pm - 12.50pm UK time, **Collocations - #LancsBox GraphColl**

Wednesday 13 November 12.00pm - 12.50pm UK time, **Group comparison – Text tool**

Wednesday 27 November 12.00pm - 12.50pm UK time, **Wordlists and keywords - Words**

Wednesday 11 December 12.00pm - 12.50pm UK time, **R scripts and #LancsBox Wizard**

CASS

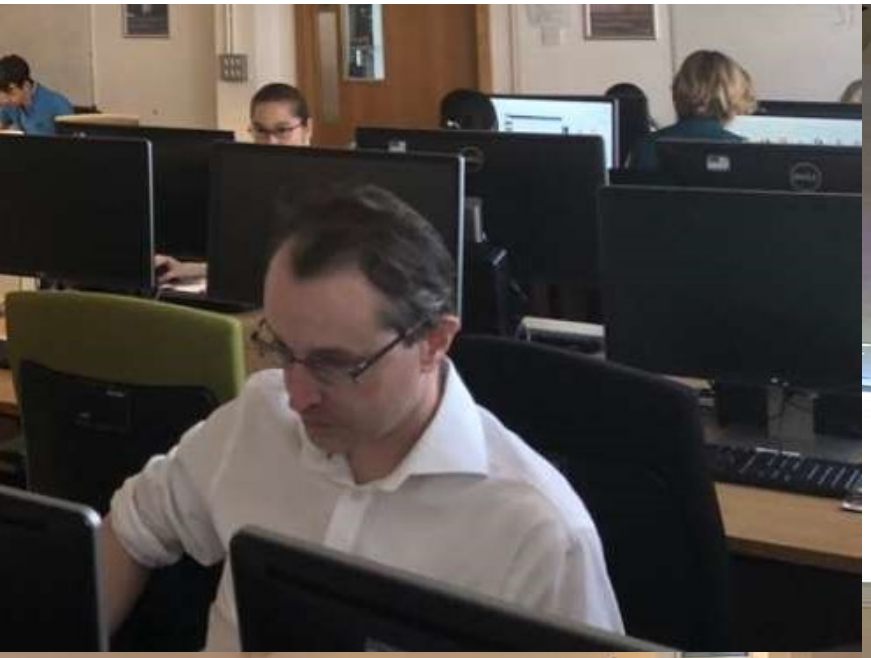
ESRC Centre for Corpus Approaches to Social Science

CORPUS APPROACH

The ESRC Centre for Corpus Approaches to Social Science (CASS) brings methodological innovation through the study of large amounts of language (corpora) with applications across a range of social sciences.

ABOUT US

https://cass.lancs.ac.uk




Effect size: Cohen's d
Cohen's d = Mean of group 1 − Mean of group 2 / pooled SD
pooled SD = √(SD1² × (cases in group1 − 1) + SD2² × (cases in group2 − 1) / all cases − 2)
The Lancaster MBA

External funding

£18.5m

People trained

75K

Publications

602

3rd

In the world for Linguistics,
QS World Subject Rankings 2024

Innovation in Corpus Linguistics.

People Online

Topics

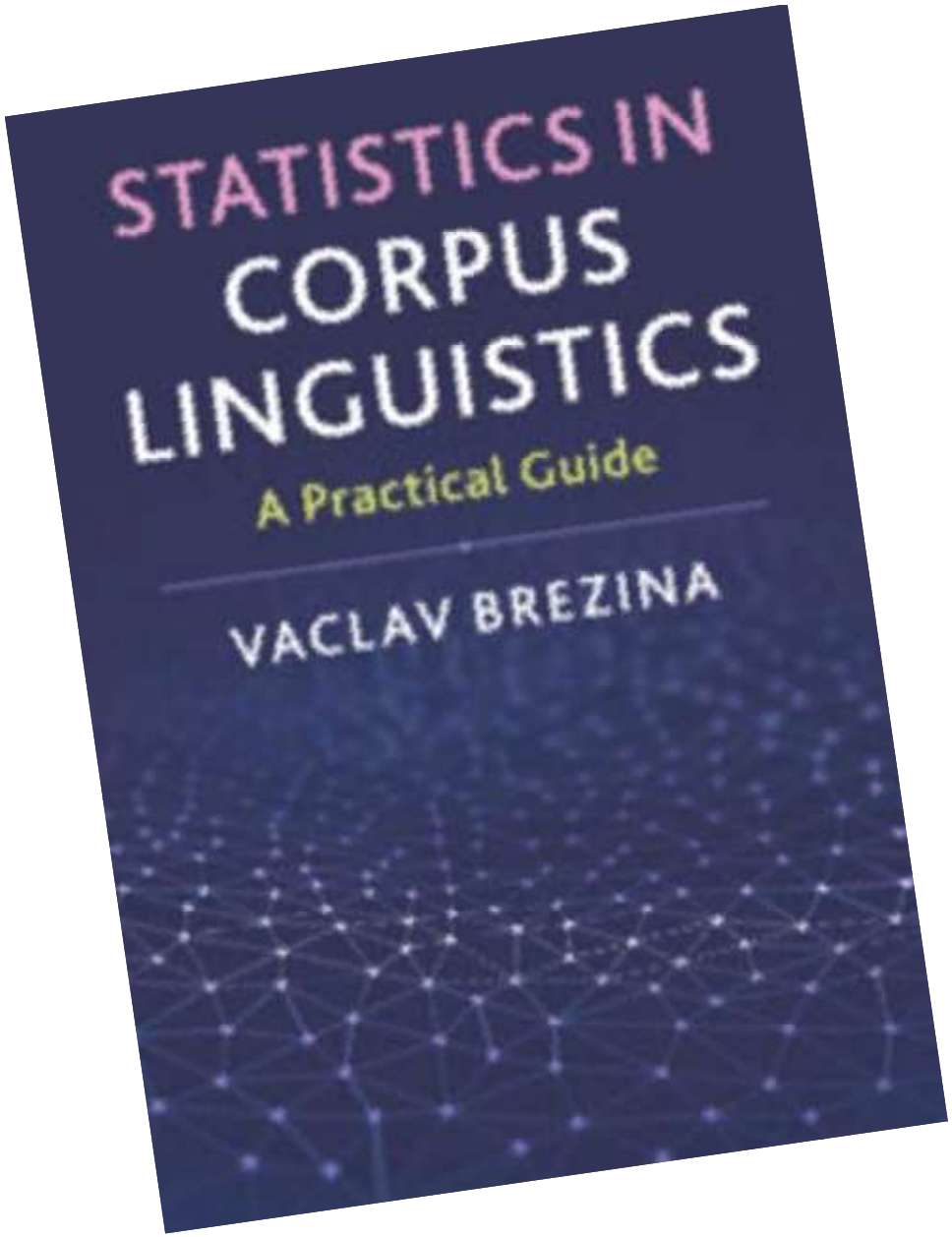
Statistics and language analysis, #LancsBox KWIC

Wednesday 30 October 12.00pm - 12.50pm UK time, **Collocations - #LancsBox GraphColl**

Wednesday 13 November 12.00pm - 12.50pm UK time, **Group comparison – Text tool**

Wednesday 27 November 12.00pm - 12.50pm UK time, **Wordlists and keywords - Words**

Wednesday 11 December 12.00pm - 12.50pm UK time, **R scripts and #LancsBox Wizard**

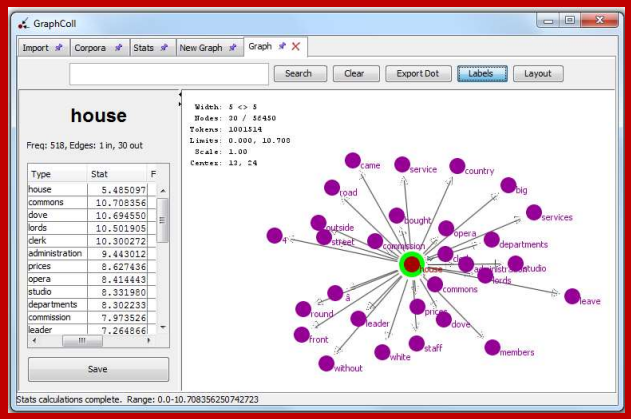


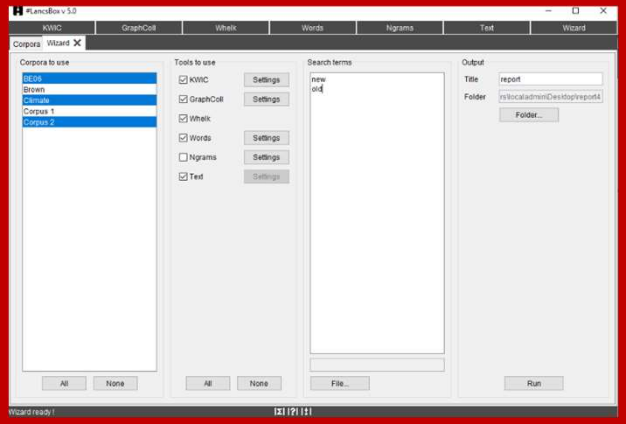
Brezina (2018)

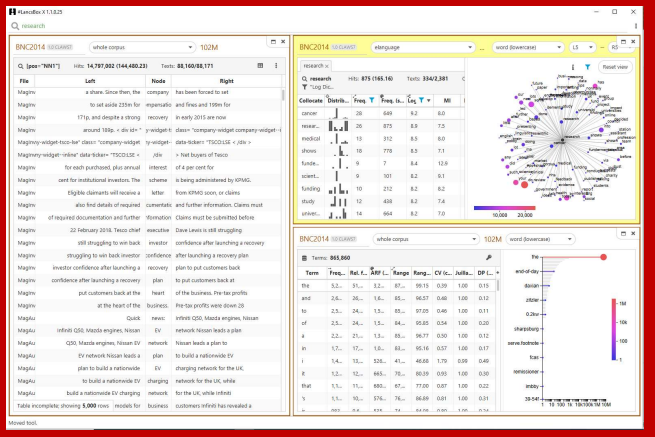
VISUALIZATION

AUTOMATION

FLEXIBILITY







2015

2020

2023

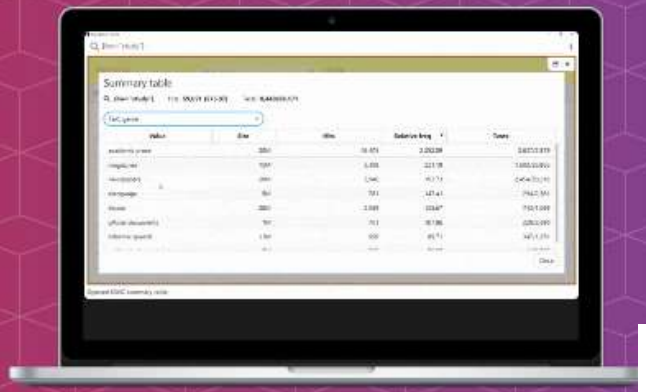
#LancsBox X

lancsbox.lancs.ac.uk

#LancsBox

DOWNLOAD FEATURES CONTACT

#LancsBox X is a powerful tool for the analysis of language: millions and billions of words.



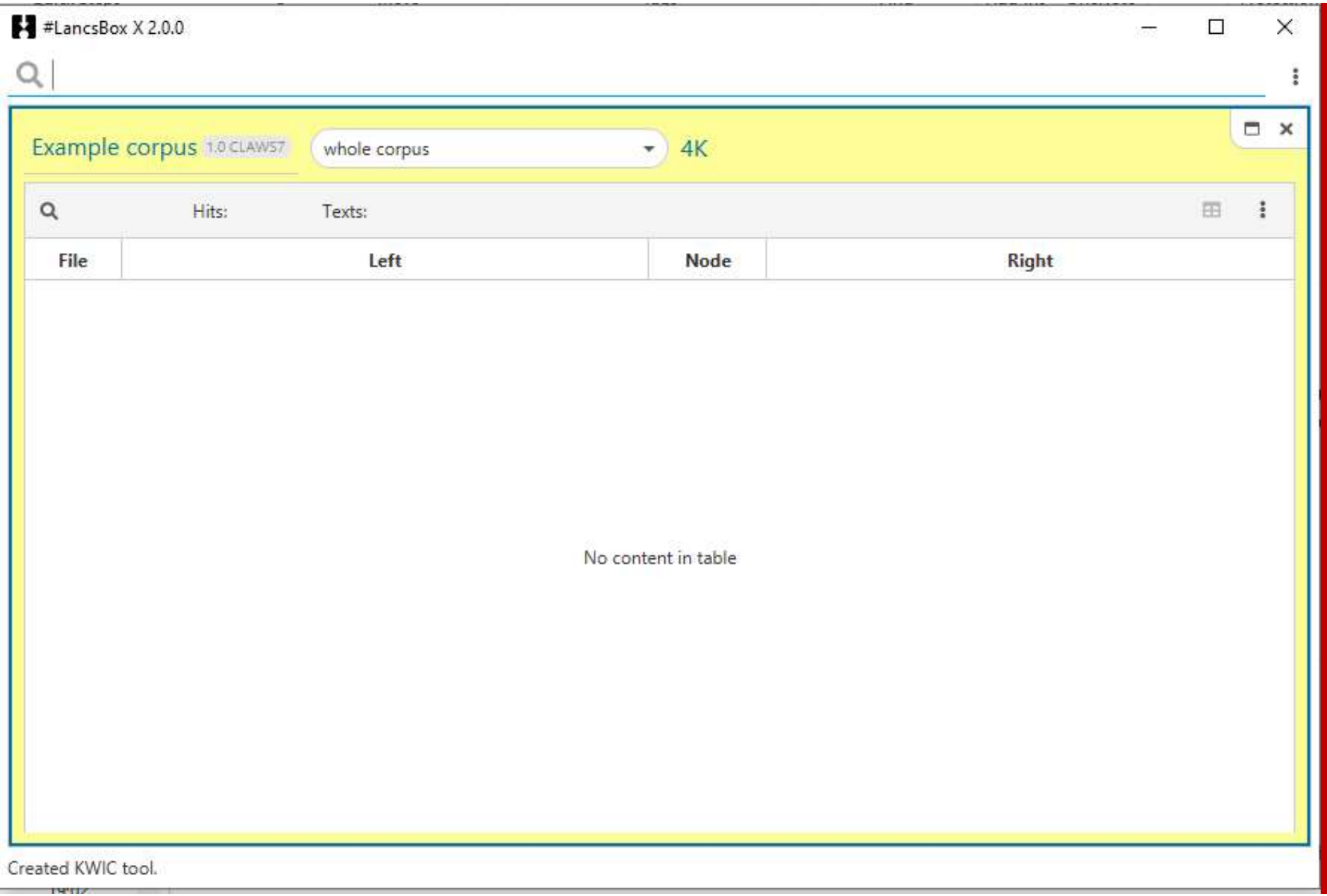
Brezina, V. & Platt, W. (2023) *#LancsBox X* [software], Lancaster University, http://lancsbox.lancs.ac.uk.

DOWNLOAD NOW FOR FREE

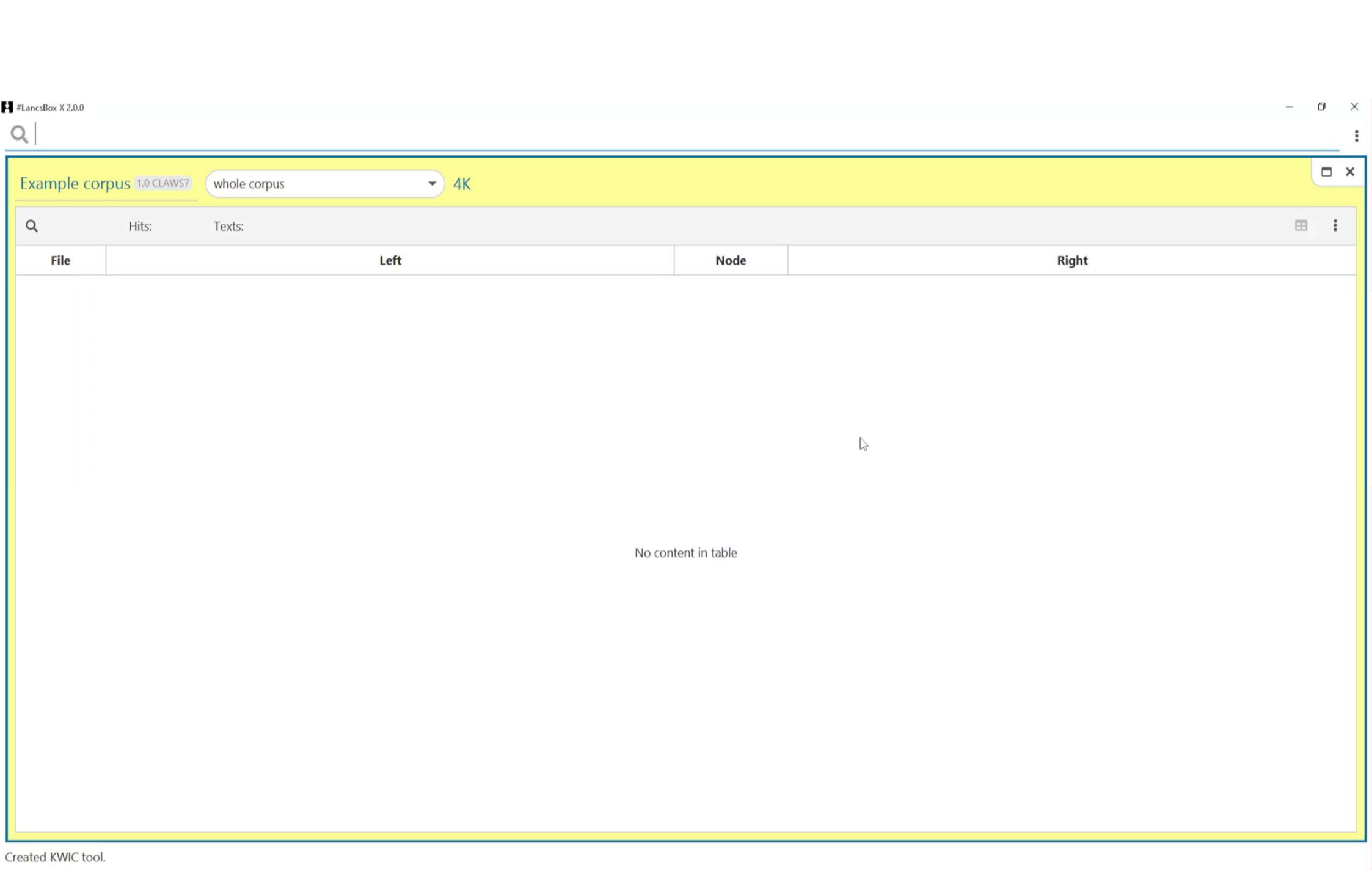
DOWNLOAD

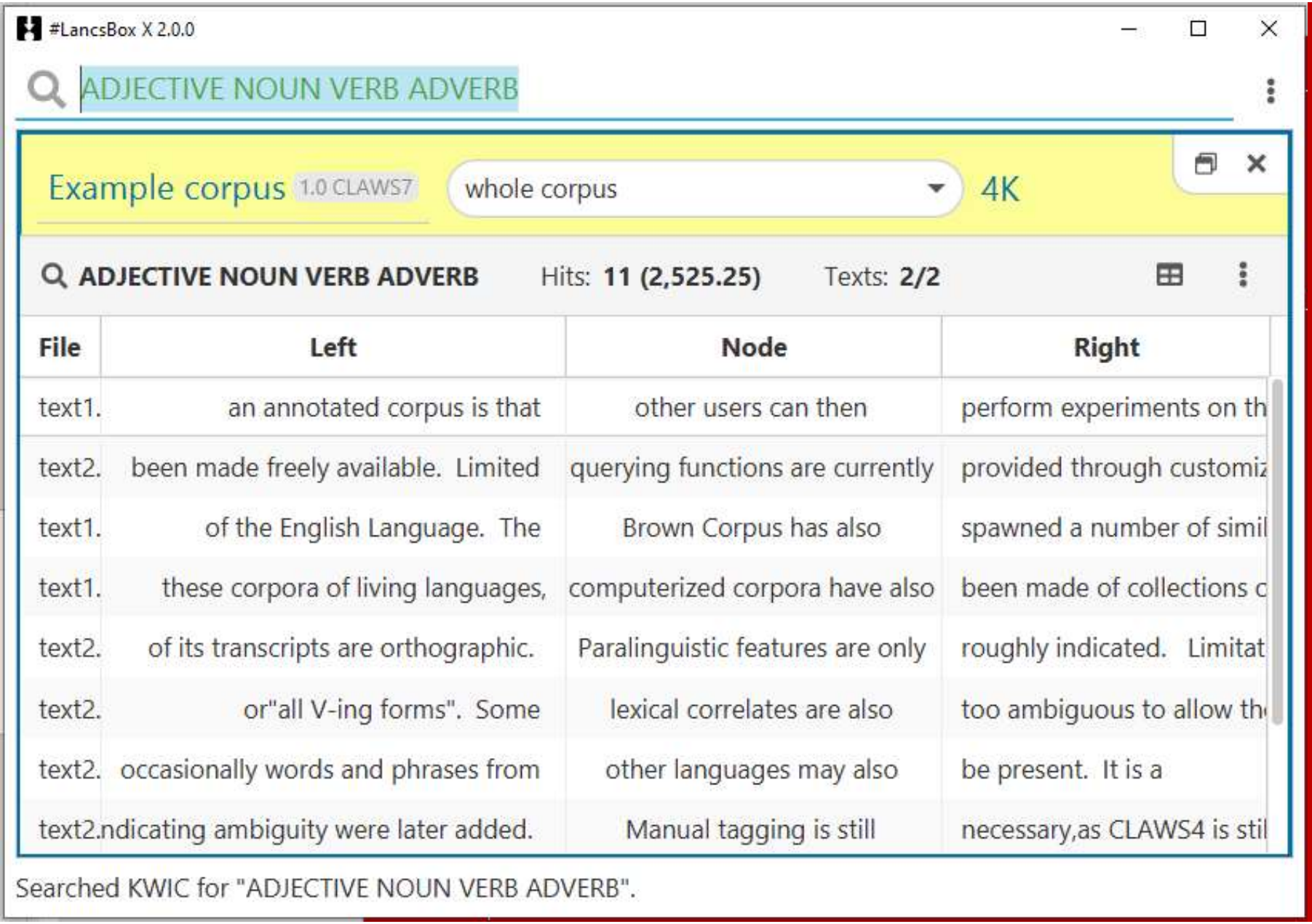
```
<text id="text1" mode="writing" genre="encyclopedia article" source="Wikipedia" url="
https://en.wikipedia.org/wiki/Corpus_linguistics" date="2023-02-21">
<s>
<w pos="NN1" hw="corpus" class="SUBST" usas="Q3">Corpus</w> <w pos="NN1" hw="linguistics" class="SUBST" usas
="Q3">linguistics</w> <w pos="VBZ" hw="be" class="VERB" usas="A3">is</w> <w pos="AT" hw="the" class="ART"
usas="Z5">the</w> <w pos="NN1" hw="study" class="SUBST" usas="P1">study</w> <w pos="IO" hw="of" class="PREP"
 usas="Z5">of</w> <w pos="AT1" hw="a" class="ART" usas="Z5">a</w> <w pos="NN1" hw="language" class="SUBST"
usas="Q3">language</w> <w pos="CSA" hw="as" class="CONJ" usas="Z5">as</w> <w pos="DD1" hw="that" class="ADJ"
 usas="Z5">that</w> <w pos="NN1" hw="language" class="SUBST" usas="Q3">language</w> <w pos="VBZ" hw="be"
class="VERB" usas="Z5">is</w> <w pos="VVN" hw="express" class="VERB" usas="Q1:1">expressed</w> <w pos="II"
hw="in" class="PREP" usas="Z5">in</w> <w pos="APPGE" hw="its" class="PRON" usas="Z8">its</w> <w pos="NN1" hw
="text" class="SUBST" usas="Q1:2">text</w> <w pos="NN1" hw="corpus" class="SUBST" usas="Q3">corpus</w><c>(
</c><w pos="JJ" hw="plural" class="ADJ" usas="Q3">plural</w> <w pos="NN2" hw="corpora" class="SUBST" usas=
"Q1:2">corpora</w><c>)</c><c>,</c><w pos="APPGE" hw="its" class="PRON" usas="Z8">its</w> <w pos="NN1" hw=
"body" class="SUBST" usas="B1">body</w> <w pos="IO" hw="of" class="PREP" usas="Z5">of</w><c>"</c><w pos="JJ"
 hw="real" class="ADJ" usas="A3">real</w> <w pos="NN1" hw="world" class="SUBST" usas="W1">world</w><c>"
</c><w pos="NN1" hw="text" class="SUBST" usas="Q1:2">text</w><c>.</c></s>
<s>
<w pos="NN1" hw="corpus" class="SUBST" usas="Q3">Corpus</w> <w pos="NN1" hw="linguistics" class="SUBST" usas
="Q3">linguistics</w> <w pos="VVZ" hw="propose" class="VERB" usas="Q2:2">proposes</w> <w pos="CST" hw="that"
 class="CONJ" usas="Z5">that</w> <w pos="AT1" hw="a" class="ART" usas="Z5">a</w> <w pos="JJ" hw="reliable"
class="ADJ" usas="A5:1">reliable</w> <w pos="NN1" hw="analysis" class="SUBST" usas="X2:4">analysis</w> <w
pos="IO" hw="of" class="PREP" usas="Z5">of</w> <w pos="AT1" hw="a" class="ART" usas="Z5">a</w> <w pos="NN1"
hw="language" class="SUBST" usas="Q3">language</w> <w pos="VBZ" hw="be" class="VERB" usas="A3">is</w> <w pos
="RGR" hw="more" class="ADV" usas="A13:3">more</w> <w pos="JJ" hw="feasible" class="ADJ" usas="A7">feasible
</w> <w pos="IW" hw="with" class="PREP" usas="Z5">with</w> <w pos="NN2" hw="corpora" class="SUBST" usas=
"Q1:2">corpora</w> <w pos="VVN" hw="collect" class="VERB" usas="A9">collected</w> <w pos="II" hw="in" class=
"PREP" usas="Z5">in</w> <w pos="AT" hw="the" class="ART" usas="Z5">the</w> <w pos="NN1" hw="fieldthe" class=
"SUBST" usas="Z99">fieldthe</w> <w pos="JJ" hw="natural" class="ADJ" usas="A6:2">natural</w> <w pos="NN1" hw
="context" class="SUBST" usas="O4:1">context</w><c>(</c><c>"</c><w pos="NN1" hw="realia" class="SUBST" usas=
"Z99">realia</w><c>"</c><c>)</c><w pos="IO" hw="of" class="PREP" usas="Z5">of</w> <w pos="DD1" hw="that"
class="ADJ" usas="Z5">that</w> <w pos="NN1" hw="languagewith" class="SUBST" usas="Z99">languagewith</w> <w
pos="JJ" hw="minimal" class="ADJ" usas="N5">minimal</w> <w pos="JJ" hw="experimental" class="ADJ" usas=
"X2:4">experimental</w> <w pos="NN1" hw="interference" class="SUBST" usas="S1:1:3">interference</w><c>.
</c></s>
```

XML


#LancsBox X 2.0.0
Example corpus 1.0 CLAWS7
whole corpus
4K
Hits:
Texts:
File
Left
Node
Right
No content in table
Created KWIC tool.

START


#LancsBox X 2.0.0
Example corpus
1.0 CLAWS7
whole corpus
4K
Hits:
Texts:
File
Left
Node
Right
No content in table
Created KWIC tool.


#LancsBox X 2.0.0
ADJECTIVE NOUN VERB ADVERB
Example corpus 1.0 CLAWS7
whole corpus
4K
ADJECTIVE NOUN VERB ADVERB Hits: 11 (2,525.25) Texts: 2/2
File Left Node Right
text1. an annotated corpus is that other users can then perform experiments on th
text2. been made freely available. Limited querying functions are currently provided through customiz
text1. of the English Language. The Brown Corpus has also spawned a number of simil
text1. these corpora of living languages, computerized corpora have also been made of collections c
text2. of its transcripts are orthographic. Paralinguistic features are only roughly indicated. Limitat
text2. or"all V-ing forms". Some lexical correlates are also too ambiguous to allow th
text2. occasionally words and phrases from other languages may also be present. It is a
text2.ndicating ambiguity were later added. Manual tagging is still necessary,as CLAWS4 is stil
Searched KWIC for "ADJECTIVE NOUN VERB ADVERB".

SEARCH

#LancsBox X 2.0.0

now

Example corpus 1.0 CLAWS7 | whole corpus | 4K

now Hits: 4 (918.27) Texts: 2/2

File	Left	Node	Right
text2.x	problem. Besides domain,there are	now	70 categories for genre for
text2.x	data,and so researchers can	now	specifically retrieve texts by genre.
text1.x	derived from source texts,but	now	that work is automated. Corpora
text1.x	Contemporary American English(1990present)is	now	available through a web interface.

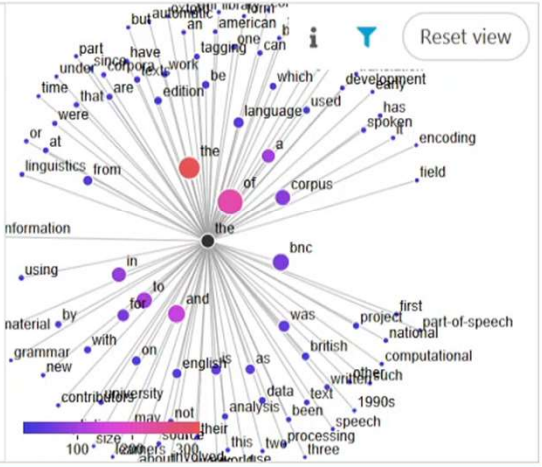
Example c... 1.0 CLAWS7 | whole corpus | ... | word (lowerca...

the ×

the Hits: 322 (73,921.03) Texts: 2/2

"Log Dic...

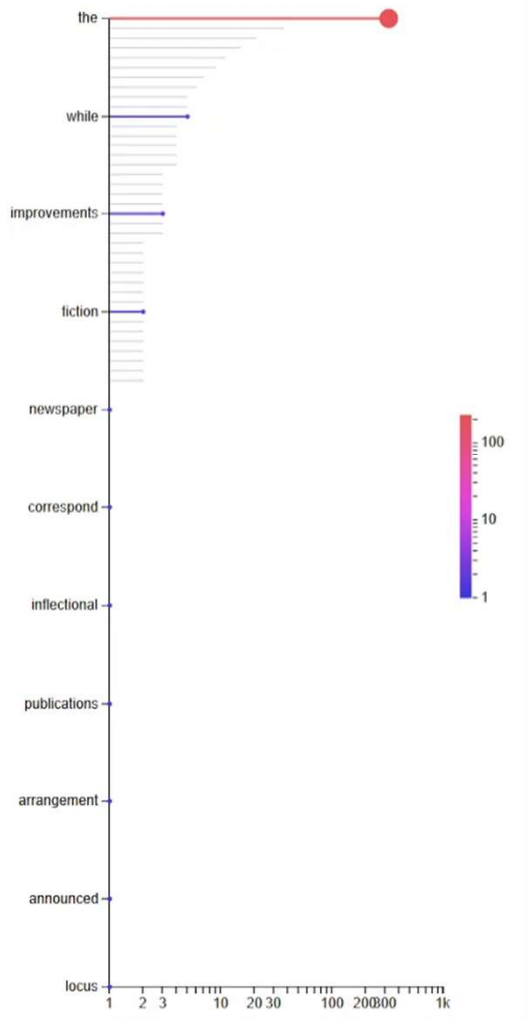
Colloc...	Distrib...	Frec	Freq. (...	Lc	MI
of		220	219	13.7	3.8
bnc		92	66	12.9	4.2
the		152	322	12.9	2.7
and		102	140	12.8	3.3
corpus		78	78	12.6	3.8
to		83	113	12.6	3.3



Example corpus 1.0 CLAWS7 | whole corpus | 4K | word (lowercase)

Terms: 1,265

Term	Fre...	Rel...	AR...	Ra...	Ra...	CV ...	Jui...	DP ...
the	3...	7...	2...	2	1...	0...	0...	0...
of	2...	5...	1...	2	1...	0...	0...	0...
a...	1...	3...	9...	2	1...	0...	0...	0...
to	1...	2...	7...	2	1...	0...	0...	0...
a	1...	2...	6...	2	1...	0...	0...	0...
in	91	2...	5...	2	1...	0...	0...	0...
c...	78	1...	4...	2	1...	0...	0...	0...
b...	66	1...	3...	1	5...	1...	0	0...
for	64	1...	4...	2	1...	0...	0...	0...
...	46	1...	2...	2	1...	0...	0...	0...
is	46	1...	2...	2	1...	0...	0...	0...
as	43	9...	2...	2	1...	0...	0...	0...
e...	41	9...	1...	2	1...	0...	0...	0...
I...	37	8...	1...	2	1...	0...	0...	0...
fr...	35	8...	2...	2	1...	0...	0...	0...
be	29	6...	1...	2	1...	0...	0...	0...
...	27	6...	1...	2	1...	0...	0...	0...
t...	27	6...	1...	2	1...	0...	0...	0...
by	26	5...	1...	2	1...	0...	0...	0...



Restored tool.

#LancsBox X 2.0.0

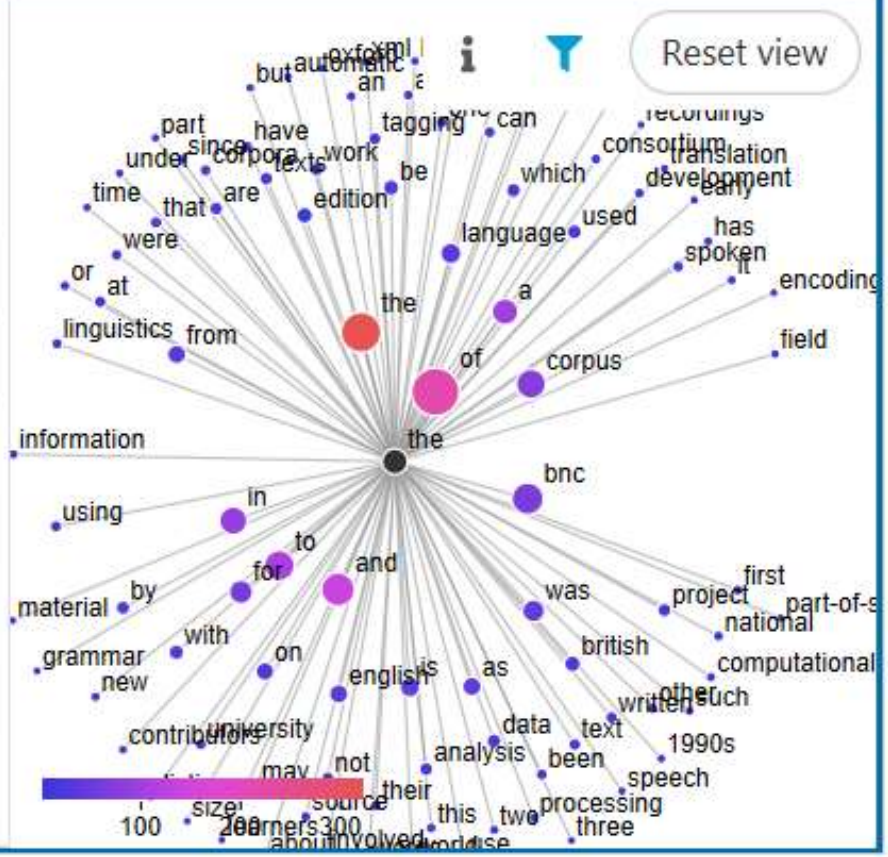
Exampl... 1.0 CLAWS7 | whole corpus | ... | word (lo...

the ×

the Hits: **322 (73,921.03)** Texts:

"Log Dic...

Colloc...	Distrib...	Frec	Freq. (...	Lc
of		220	219	13.7
bnc		92	66	12.9
the		152	322	12.9
and		102	140	12.8
corpus		78	78	12.6
to		83	113	12.6



Maximized tool.

GRAPH

#LancsBox X 2.0.0

data

Example corpus 1.0 CLAWS7 | whole corpus | 4K

[hw="p.*" pos="n.*"] Hits: 91 (20,890.73) Texts: 2/2

File	Left	Node	Right
text1.x	science,containing relevant computer science	publications	with sentient metadata such as
text1.x	NLP Scholar,a combination of	papers	of the ACL Anthology and
text1.x	which attempt to trace a	path	from data to theory. Wallis
text1.x	what they called the 3A	perspective:	Annotation,Abstraction and Analysis. Ar
text1.x	may include e.g.,rule-learning for	parsers.	Analysis consists of statistically probing,
text1.x	lexical corpora today are part-of-speech-tagged(	POS-tagged).	However even corpus linguists who
text1.x	with other interests and differing	perspectives	than the originators'can exploit

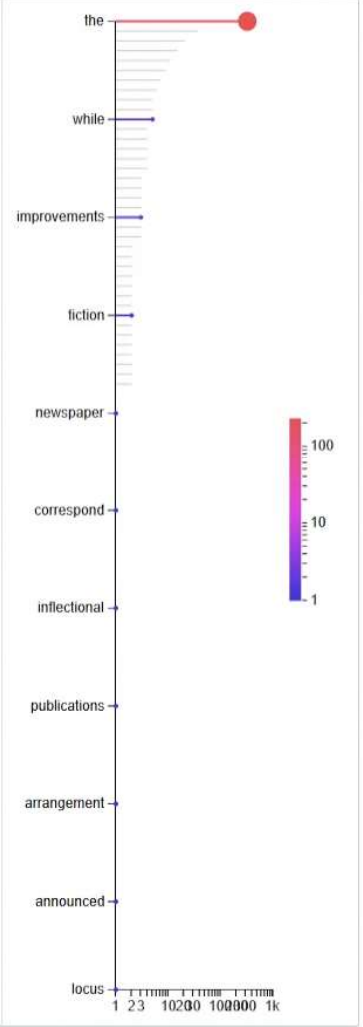
Example c... 1.0 CLAWS7 | whole corpus | ... | word (lowerca...

Reset view

Example corpus 1.0 CLAWS7 | whole corpus | 4K | word (lowercase)

Terms: 1,265

Term	Freq...	Rel. f...	ARF (...	Range	Rang...	CV (c...	Juilla...	DP (d...
the	322	73,...	225...	2	100...	0.11	0.89	0.04
of	219	50,...	145...	2	100...	0.13	0.87	0.06
and	140	32,...	96.12	2	100...	0.08	0.92	0.03
to	113	25,...	72.62	2	100...	0.26	0.74	0.09
a	102	23,...	66.51	2	100...	0.12	0.88	0.05
in	91	20,...	59.43	2	100...	0.14	0.86	0.06
cor...	78	17,...	42.26	2	100...	0.35	0.65	0.17
bnc	66	15,...	34.47	1	50.00	1.00	0	0.28
for	64	14,...	40.05	2	100...	0.26	0.74	0.09
was	46	10,...	23.85	2	100...	0.17	0.83	0.06
is	46	10,...	24.13	2	100...	0.05	0.95	0.02
as	43	9,8...	28.45	2	100...	0.06	0.94	0.02
eng...	41	9,4...	16.63	2	100...	0.50	0.50	0.26
lan...	37	8,4...	19.84	2	100...	0.37	0.63	0.18
from	35	8,0...	21.80	2	100...	0.31	0.69	0.11
be	29	6,6...	16.95	2	100...	0.68	0.32	0.21
with	27	6,1...	15.66	2	100...	0.12	0.88	0.05
that	27	6,1...	15.96	2	100...	0.12	0.88	0.05
by	26	5,9...	19.35	2	100...	0.03	0.97	0.01



Created GraphColl tool.

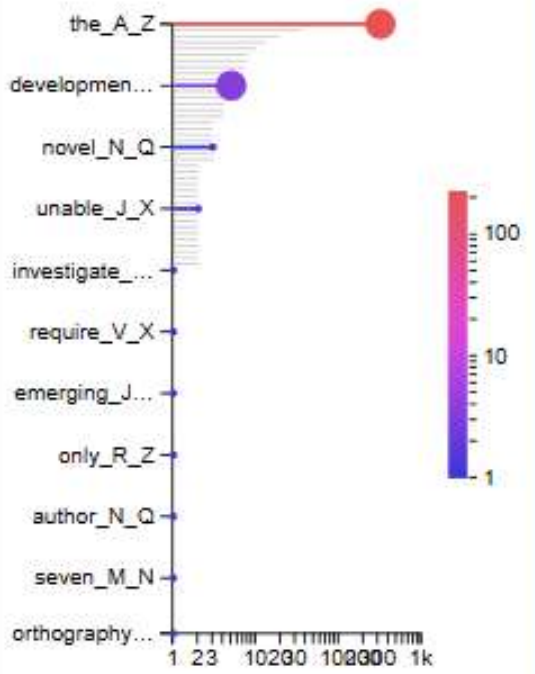
#LancsBox X 2.0.0

c*

Example cor... 1.0 CLAWS7 | whole corpus | ... | lexeme

Terms: **1,253**

Term	Freq...	Rel. ...	ARF ...	Range	Ran...	CV (...	Juill..D
right_N_S	3	68...	1.12	1	50....	1.00	0
three_M_N	3	68...	2.44	1	50....	1.00	0
range_N_N	3	68...	2.27	2	10...	0.12	0.88
now_R_T	3	68...	2.01	2	10...	0.12	0.88
creation_N_A	3	68...	2.05	1	50....	1.00	0
success_N_X	3	68...	2.25	2	10...	0.12	0.88



Closed keywords table.

WORDS

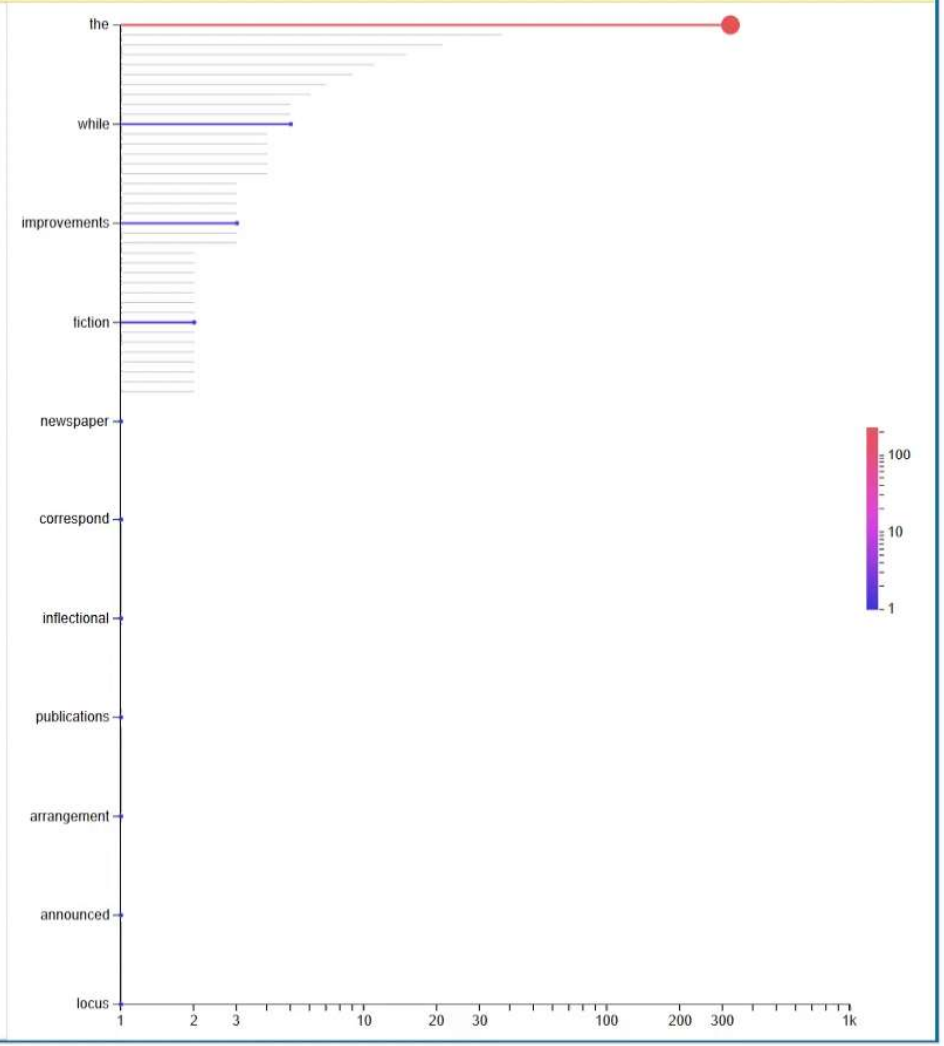
#LancsBox X 2.0.0

corpus

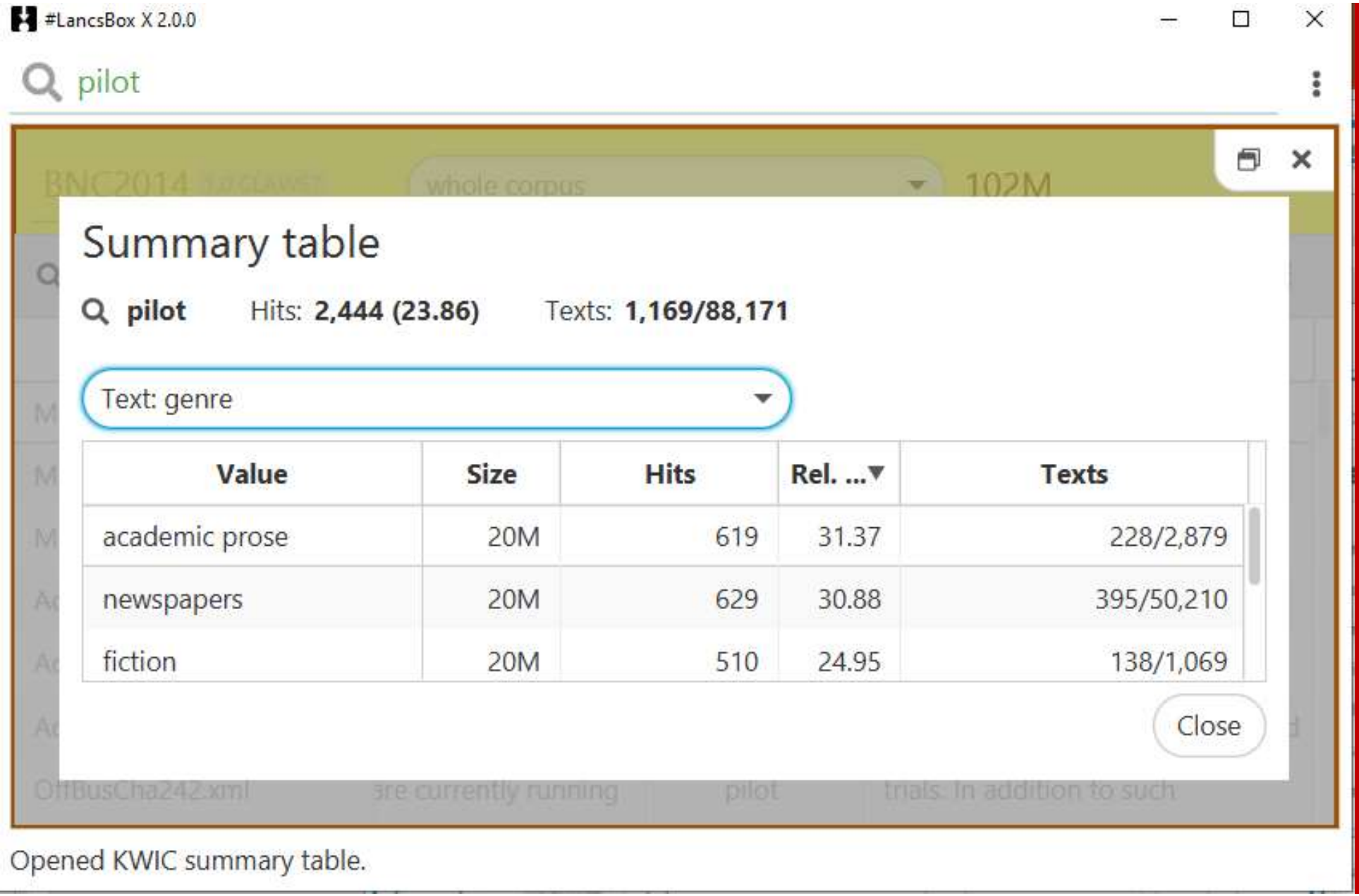
Example corpus 1.0 CLAWS7 | whole corpus | 4K | word (lowercase)

Terms: **1,265**

Term	Frequency	Rel. frequency	ARF (averag...	Range	Range %	CV (coeffici...	Juilland's D	DP (deviatio...
their	17	3,902.66	6.47	1	50.00	1.00	0	0.28
work	17	3,902.66	11.63	2	100.00	0.17	0.83	0.07
been	16	3,673.09	8.36	2	100.00	0.08	0.92	0.03
at	16	3,673.09	10.19	2	100.00	0.08	0.92	0.03
text	16	3,673.09	9.64	2	100.00	0.26	0.74	0.09
such	15	3,443.53	10.11	2	100.00	0.12	0.88	0.05
not	15	3,443.53	6.89	2	100.00	0.69	0.31	0.21
it	15	3,443.53	8.87	2	100.00	0.69	0.31	0.21
this	15	3,443.53	9.38	2	100.00	0.43	0.57	0.15
an	14	3,213.96	9.19	2	100.00	0.01	0.99	0.006
words	14	3,213.96	8.50	2	100.00	0.40	0.60	0.14
linguistics	14	3,213.96	7.15	2	100.00	0.81	0.19	0.51
can	14	3,213.96	7.75	2	100.00	0.18	0.82	0.07
have	13	2,984.39	7.59	2	100.00	0.23	0.77	0.10
also	13	2,984.39	8.65	2	100.00	0.13	0.87	0.05
one	13	2,984.39	7.22	2	100.00	0.36	0.64	0.13
tagging	13	2,984.39	3.67	2	100.00	0.36	0.64	0.13
may	12	2,754.82	7.25	2	100.00	0.32	0.68	0.11
has	11	2,525.25	6.34	2	100.00	0.19	0.81	0.08



Maximized tool.


#LancsBox X 2.0.0
pilot
Summary table
pilot Hits: 2,444 (23.86) Texts: 1,169/88,171
Text: genre
Value Size Hits Rel. ... Texts
academic prose 20M 619 31.37 228/2,879
newspapers 20M 629 30.88 395/50,210
fiction 20M 510 24.95 138/1,069
Close
Opened KWIC summary table.

SUMMARY

#LancsBox X 2.0.0

Example corpus 1.0 CLAWS7 | whole corpus | 4K

[hw="p.*" pos="n.*"] Hits: 91 (20,890.73) Texts: 2/2

File	Left	Node	Right
text1.x	science,containing relevant computer science	publications	with sentient metadata such as
text1.x	NLP Scholar,a combination of	papers	of the ACL Anthology and
text1.x	which attempt to trace a	path	from data to theory. Wallis
text1.x	what they called the 3A	perspective:	Annotation,Abstraction and Analysis. Ar
text1.x	may include e.g.,rule-learning for	parsers.	Analysis consists of statistically probing,
text1.x	lexical corpora today are part-of-speech-tagged(	POS-tagged).	However even corpus linguists who
text1.x	with other interests and differing	perspectives	than the originators'can exploit

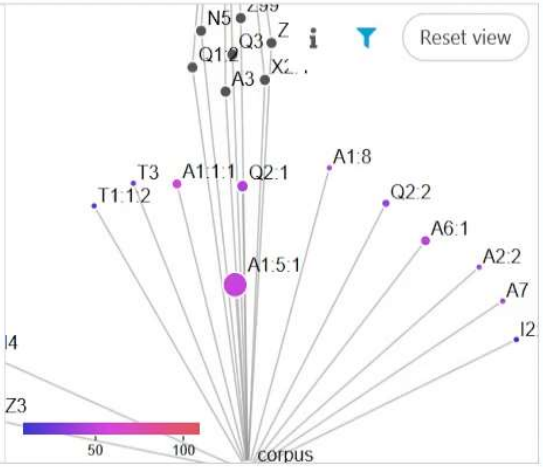
Example c... 1.0 CLAWS7 | whole corpus | ... | usas

data × corpus ×

corpus Hits: 78 (17,906.34) Texts: 2/2

"Freq. (c...

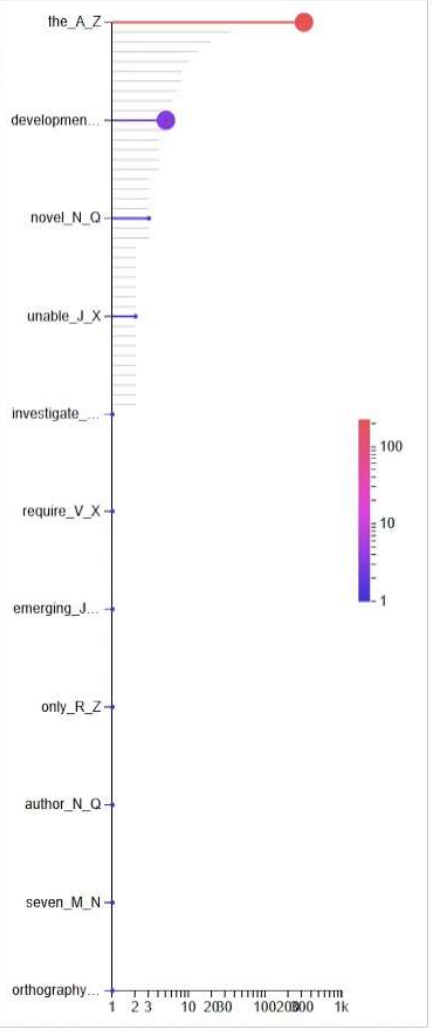
Colloc...	Distrib...	Fre	Freq. (...	Log Dice	MI
Q3		40	250	12.0	3.2
A1:1:1		7	56	10.7	2.8
A6:1		7	49	10.8	3.0
A3		21	96	11.9	3.6
A7		5	43	10.4	2.7
A2:2		5	46	10.4	2.6



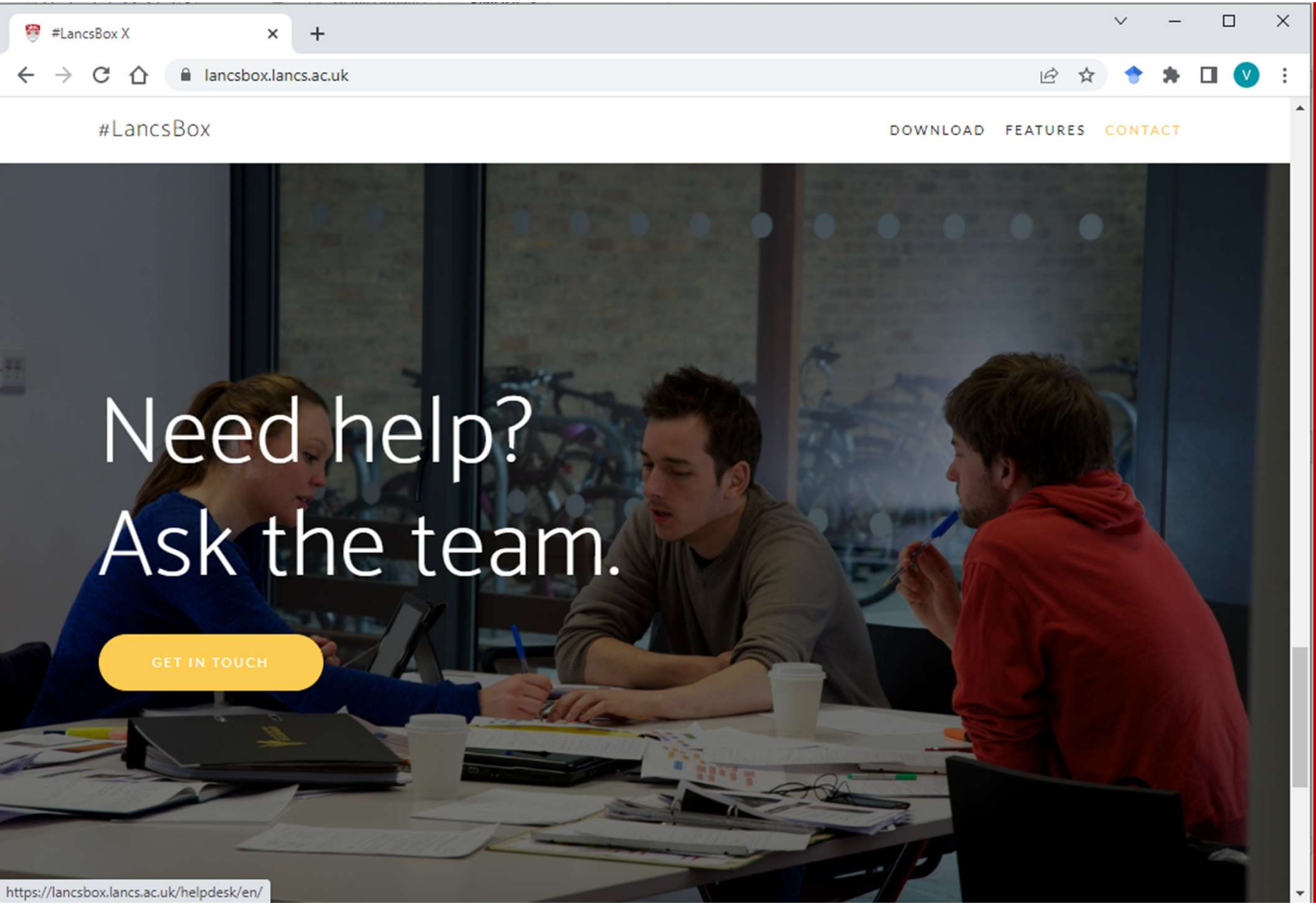
Example corpus 1.0 CLAWS7 | whole corpus | 4K | lexeme

Terms: 1,253

Term	Fre...	Rel....	AR...	Ran...	Ran...	CV ...	Juil...	DP...
incorporat...	3	6...	1....	1	5...	1....	0	...
identity_N_S	3	6...	1....	1	5...	1....	0	...
improvem...	3	6...	2....	1	5...	1....	0	...
right_N_S	3	6...	1....	1	5...	1....	0	...
three_M_N	3	6...	2....	1	5...	1....	0	...
range_N_N	3	6...	2....	2	1...	0....	0....	...
now_R_T	3	6...	2....	2	1...	0....	0....	...
creation_N...	3	6...	2....	1	5...	1....	0	...
success_N_X	3	6...	2....	2	1...	0....	0....	...
deal_V_A	3	6...	1....	1	5...	1....	0	...
mean_V_Q	3	6...	2....	1	5...	1....	0	...
tagger_N_Y	3	6...	1....	1	5...	1....	0	...
both_D_N	3	6...	1....	1	5...	1....	0	...
brown_J_Z	3	6...	1....	1	5...	1....	0	...
itself_P_Z	3	6...	1....	1	5...	1....	0	...
become_V...	3	6...	2....	1	5...	1....	0	...
who_P_Z	3	6...	1....	1	5...	1....	0	...
speak_V_Q	3	6...	2....	2	1...	0....	0....	...
a_A_N	3	6...	1....	1	5...	1....	0	...
would_V_A	3	6	1	1	5	1	0	



Restored tool.


#LancsBox X
lancsbox.lancs.ac.uk
#LancsBox
DOWNLOAD
FEATURES
CONTACT
Need help?
Ask the team.
GET IN TOUCH
https://lancsbox.lancs.ac.uk/helpdesk/en/

SUPPORT