Open research group

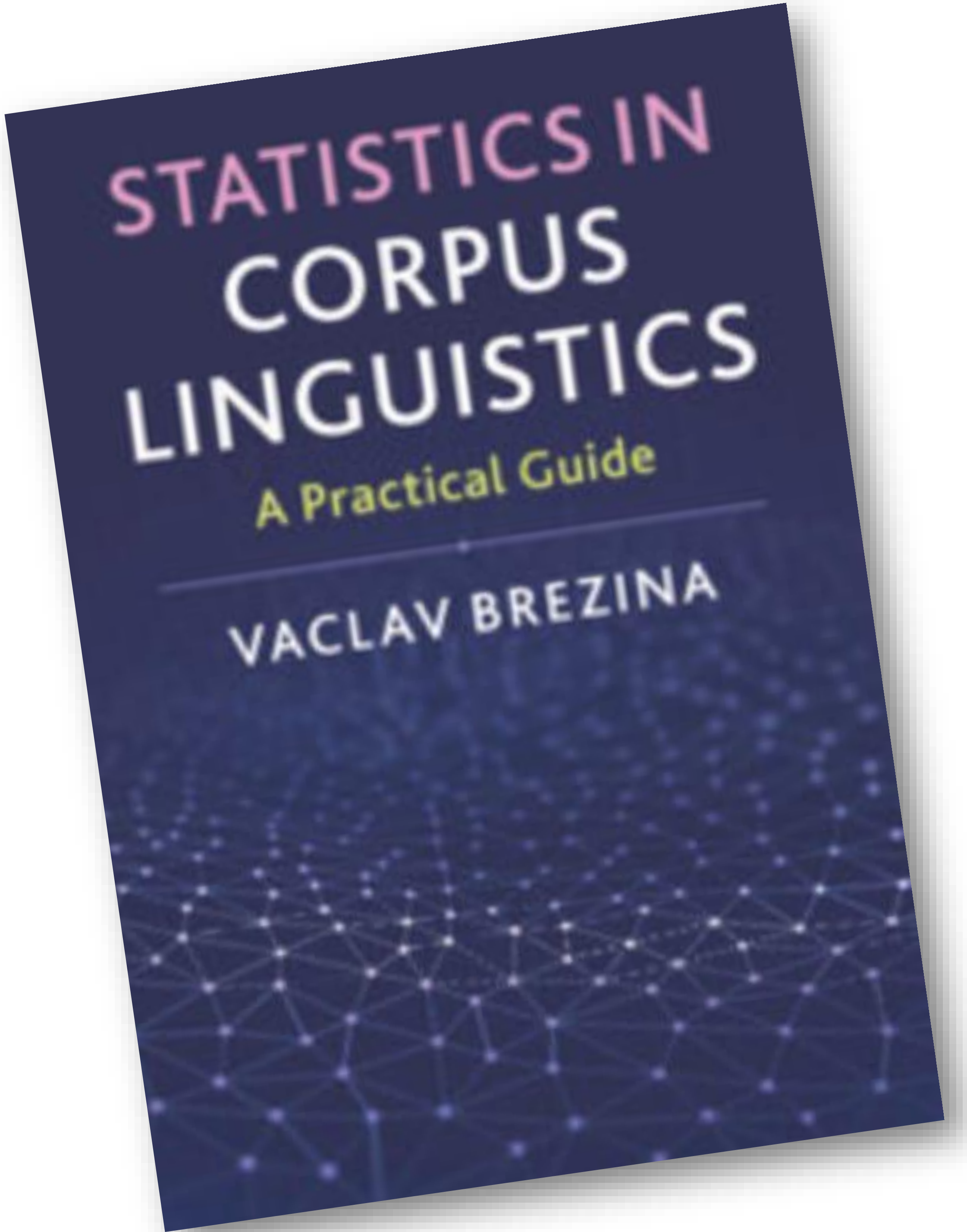
ESRC Centre for Corpus Approaches to Social Science

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**

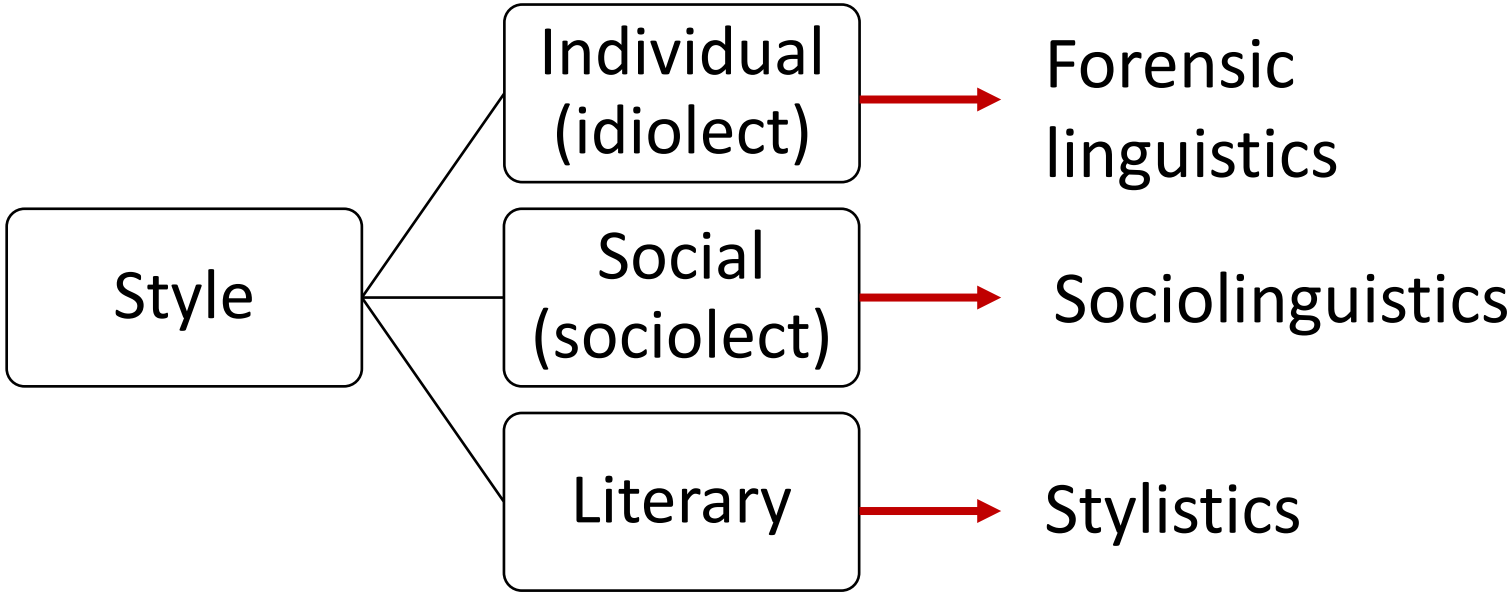


Brezina (2018)

The notion of style is central to the analyses described in this chapter. Following Coupland's (Coupland, 2007: 2) broad definition of style as 'ways of speaking that are indexically linked to social groups, times and places', we will be looking at the role of speaker background and speech community in the language that speakers produce. Style is a unifying notion linking sociolinguistics (social style), stylistics (literary style) and forensic linguistics (individual style). Regardless of whether we are looking at naturally occurring data or fiction, the statistical procedures discussed in this chapter will help us quantify and make sense of variation in speaking/writing style. Linguistic variables involved in this variation show systematic differences according to both individual speakers (distinguishing individual styles) and groups (distinguishing social dialects or sociolects). But how can we identify such variables? (pp. 184-185)

Style

- “Ways of speaking that are indexically linked to social groups, times and places” Coupland (2007: 2).



Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Think about and discuss

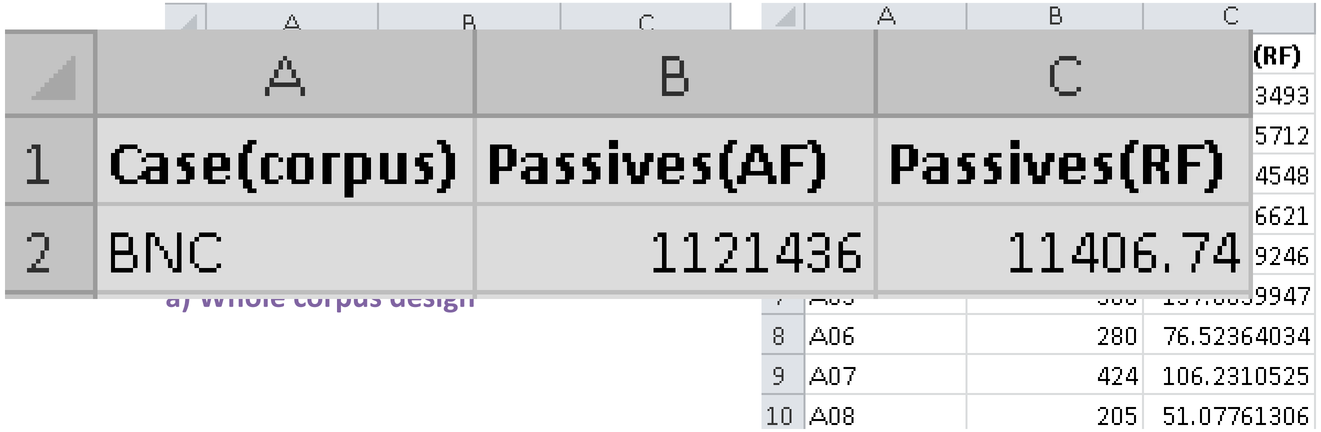
1. Does personal *style* of speaking/writing mater?
2. What is the likely effect of AI tools such as Chat GPT on personal style?
3. How can we capture/measure style statistically?

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Research design: *Where* to count linguistic features?

- Whole corpus
- Individual texts/speakers
- Linguistic features of interest

Whole corpus

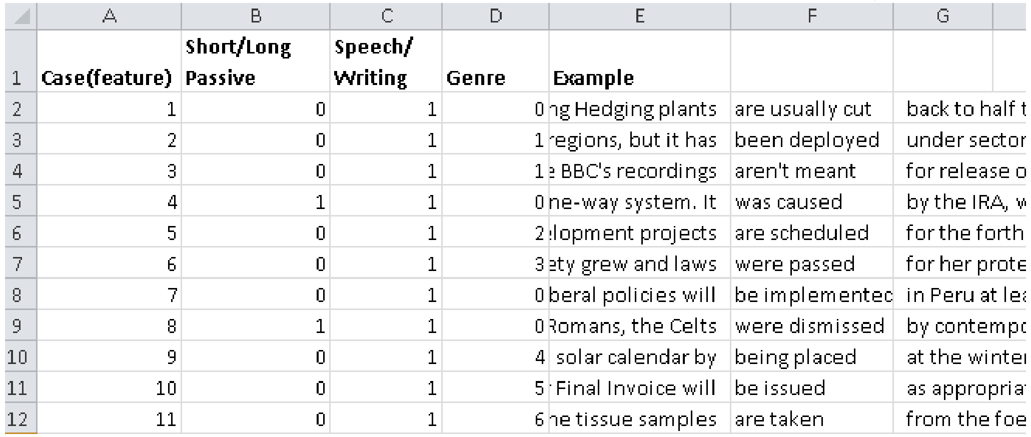


	A	B	C
1	Case(corpus)	Passives(AF)	Passives(RF)
2	BNC	1121436	11406.74

a) Whole corpus design

	A	B	C
			(RF)
			3493
			5712
			4548
			6621
			9246
7			9947
8	A06	280	76.52364034
9	A07	424	106.2310525
10	A08	205	51.07761306

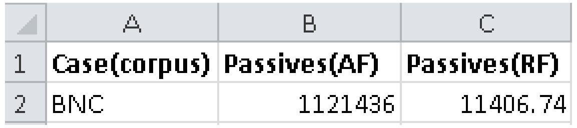
b) Individual-texts/speakers design



	A	B	C	D	E	F	G
1	Case(feature)	Short/Long Passive	Speech/ Writing	Genre	Example		
2	1	0	1	0	ng Hedging plants	are usually cut	back to half t
3	2	0	1	1	regions, but it has	been deployed	under sector
4	3	0	1	1	e BBC's recordings	aren't meant	for release o
5	4	1	1	0	ne-way system. It	was caused	by the IRA, w
6	5	0	1	2	elopment projects	are scheduled	for the forth
7	6	0	1	3	ety grew and laws	were passed	for her prote
8	7	0	1	0	beral policies will	be implementec	in Peru at lea
9	8	1	1	0	Romans, the Celts	were dismissed	by contempo
10	9	0	1	4	solar calendar by	being placed	at the winter
11	10	0	1	5	Final Invoice will	be issued	as appropria
12	11	0	1	6	he tissue samples	are taken	from the foe

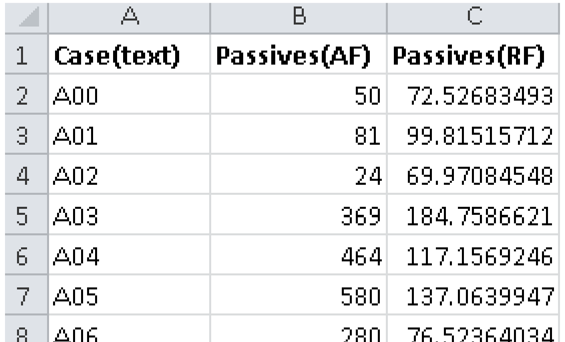
c) Linguistic feature design

Linguistic features

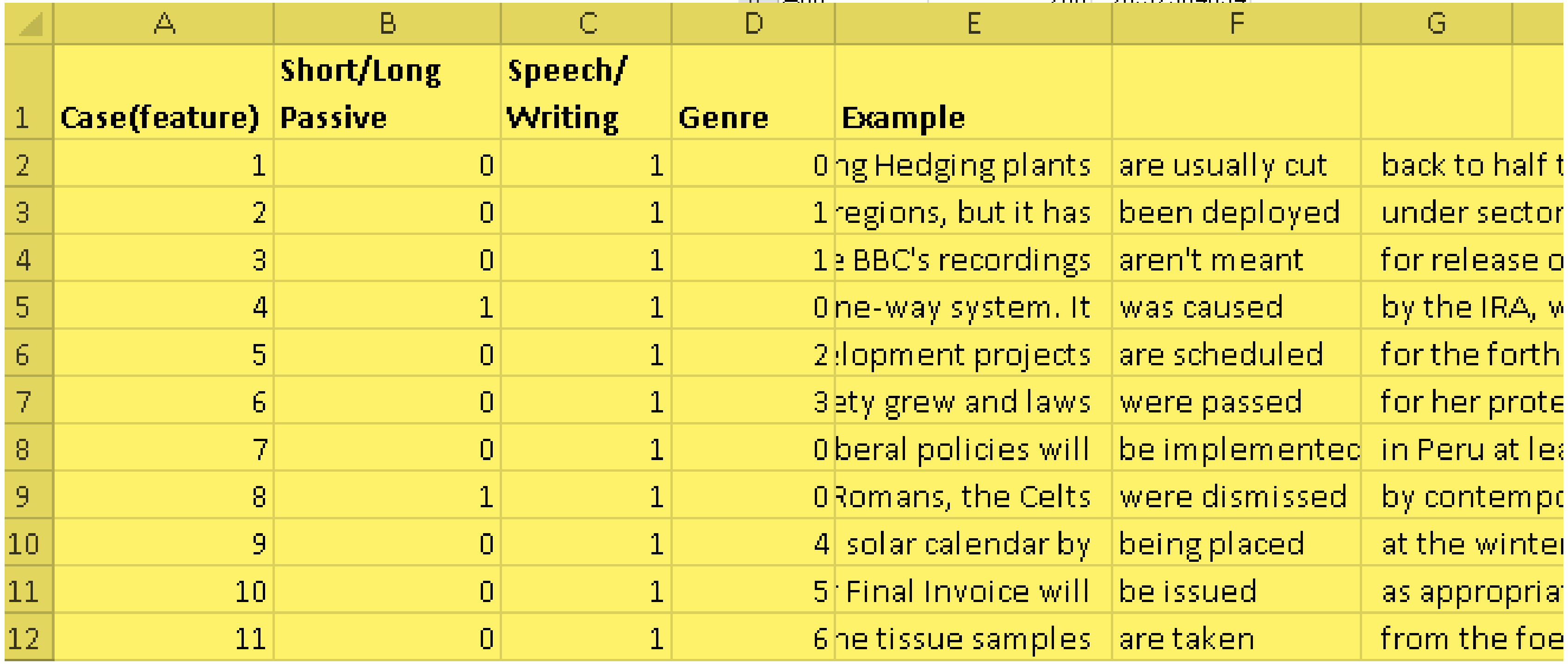


	A	B	C
1	Case(corpus)	Passives(AF)	Passives(RF)
2	BNC	1121436	11406.74

a) Whole corpus design



	A	B	C
1	Case(text)	Passives(AF)	Passives(RF)
2	A00	50	72.52683493
3	A01	81	99.81515712
4	A02	24	69.97084548
5	A03	369	184.7586621
6	A04	464	117.1569246
7	A05	580	137.0639947
8	A06	280	76.52364034

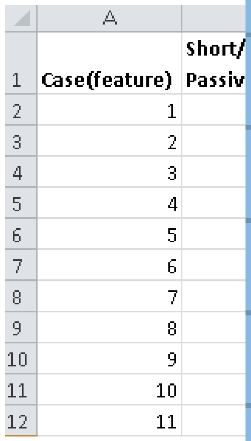


	A	B	C	D	E	F	G
1	Case(feature)	Short/Long Passive	Speech/ Writing	Genre	Example		
2	1	0	1	0	ng Hedging plants	are usually cut	back to half t
3	2	0	1	1	regions, but it has	been deployed	under sector
4	3	0	1	1	e BBC's recordings	aren't meant	for release o
5	4	1	1	0	ne-way system. It	was caused	by the IRA, w
6	5	0	1	2	lopment projects	are scheduled	for the forth
7	6	0	1	3	ety grew and laws	were passed	for her prote
8	7	0	1	0	beral policies will	be implementec	in Peru at lea
9	8	1	1	0	Romans, the Celts	were dismissed	by contempo
10	9	0	1	4	solar calendar by	being placed	at the winter
11	10	0	1	5	Final Invoice will	be issued	as appropria
12	11	0	1	6	ne tissue samples	are taken	from the foe

Individual texts/speakers

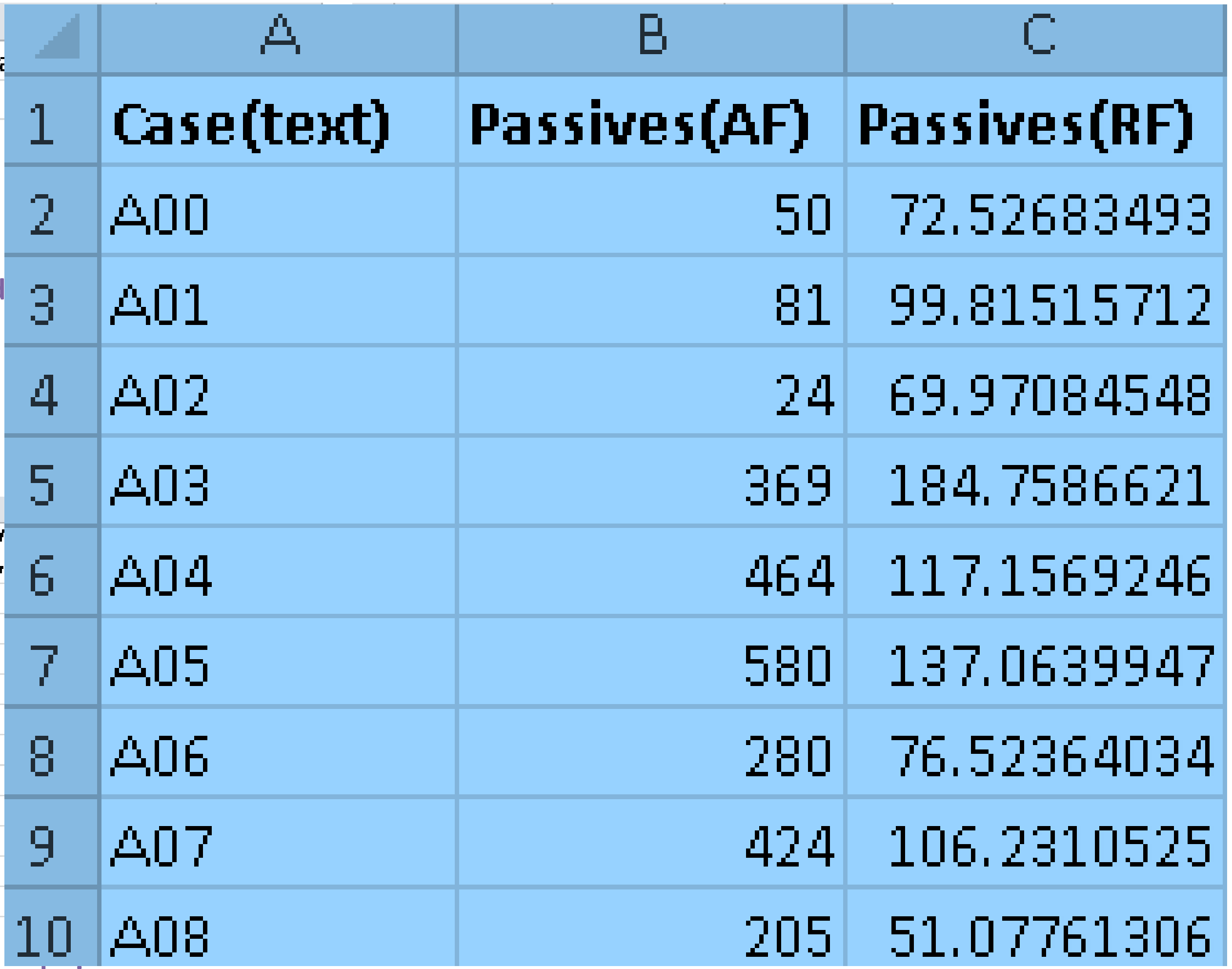
	A	
1	Case(corpus)	Pa
2	BNC	

a) Whole corpus d

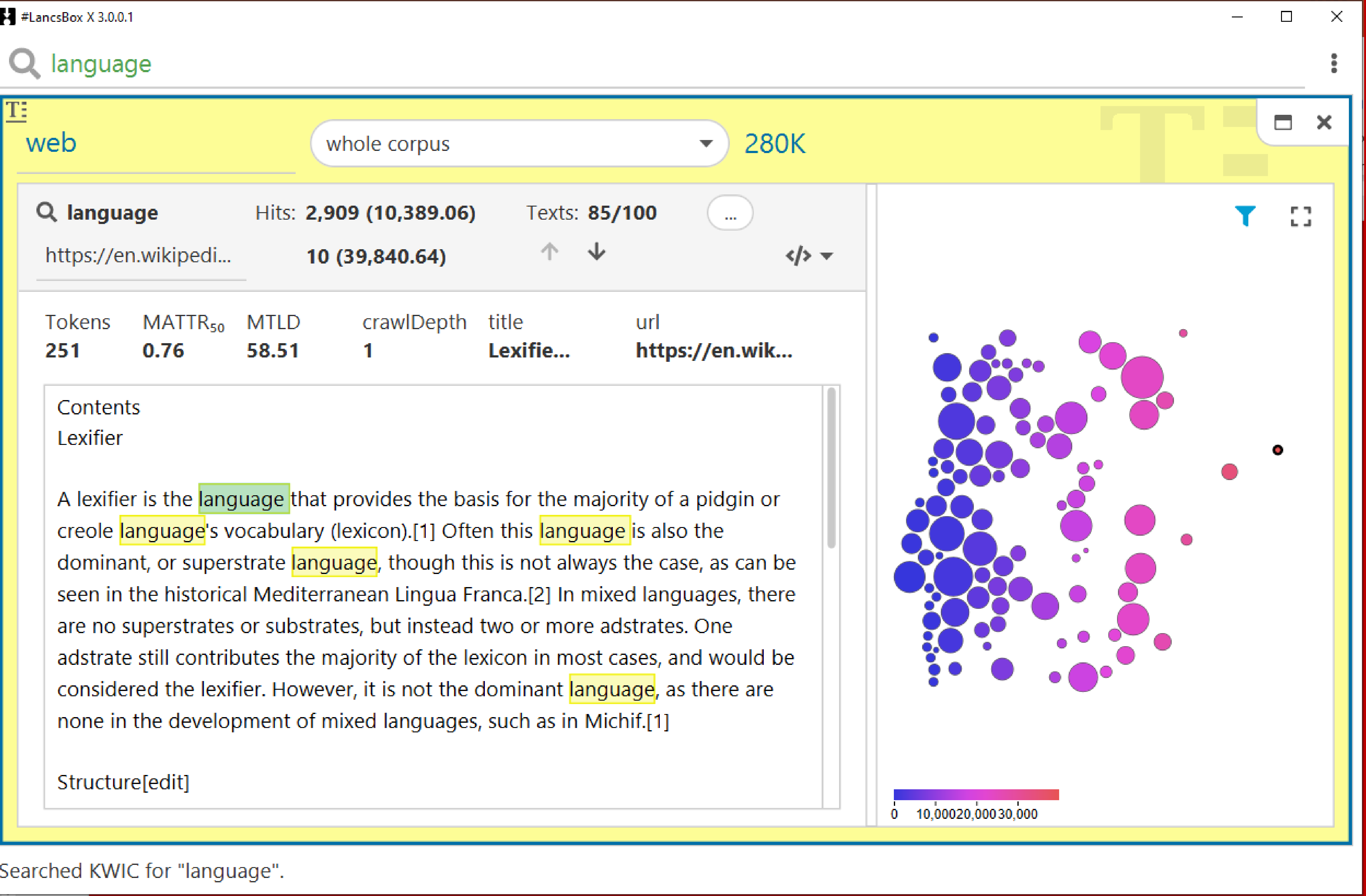


	A	
1	Case(feature)	Short/ Passiv
2	1	
3	2	
4	3	
5	4	
6	5	
7	6	
8	7	
9	8	
10	9	
11	10	
12	11	

c) Linguistic feature design



	A	B	C
1	Case(text)	Passives(AF)	Passives(RF)
2	A00	50	72.52683493
3	A01	81	99.81515712
4	A02	24	69.97084548
5	A03	369	184.7586621
6	A04	464	117.1569246
7	A05	580	137.0639947
8	A06	280	76.52364034
9	A07	424	106.2310525
10	A08	205	51.07761306



Tokens	MATTR$_{50}$	MTLD	crawlDepth	title	url
251	0.76	58.51	1	Lexifie...	https://en.wik...

Contents

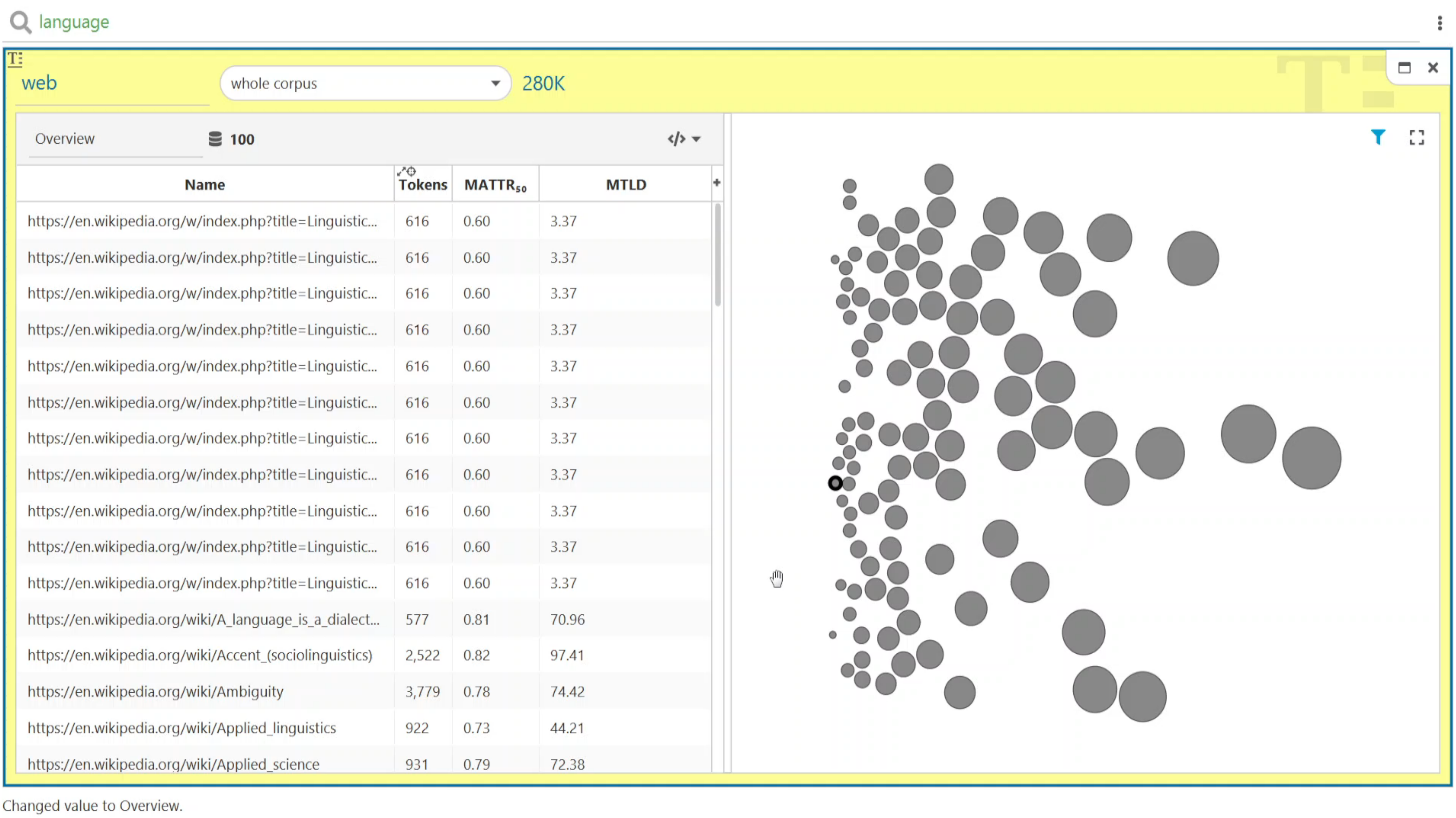
Lexifier

A lexifier is the language that provides the basis for the majority of a pidgin or creole language's vocabulary (lexicon).[1] Often this language is also the dominant, or superstrate language, though this is not always the case, as can be seen in the historical Mediterranean Lingua Franca.[2] In mixed languages, there are no superstrates or substrates, but instead two or more adstrates. One adstrate still contributes the majority of the lexicon in most cases, and would be considered the lexifier. However, it is not the dominant language, as there are none in the development of mixed languages, such as in Michif.[1]

Structure[edit]

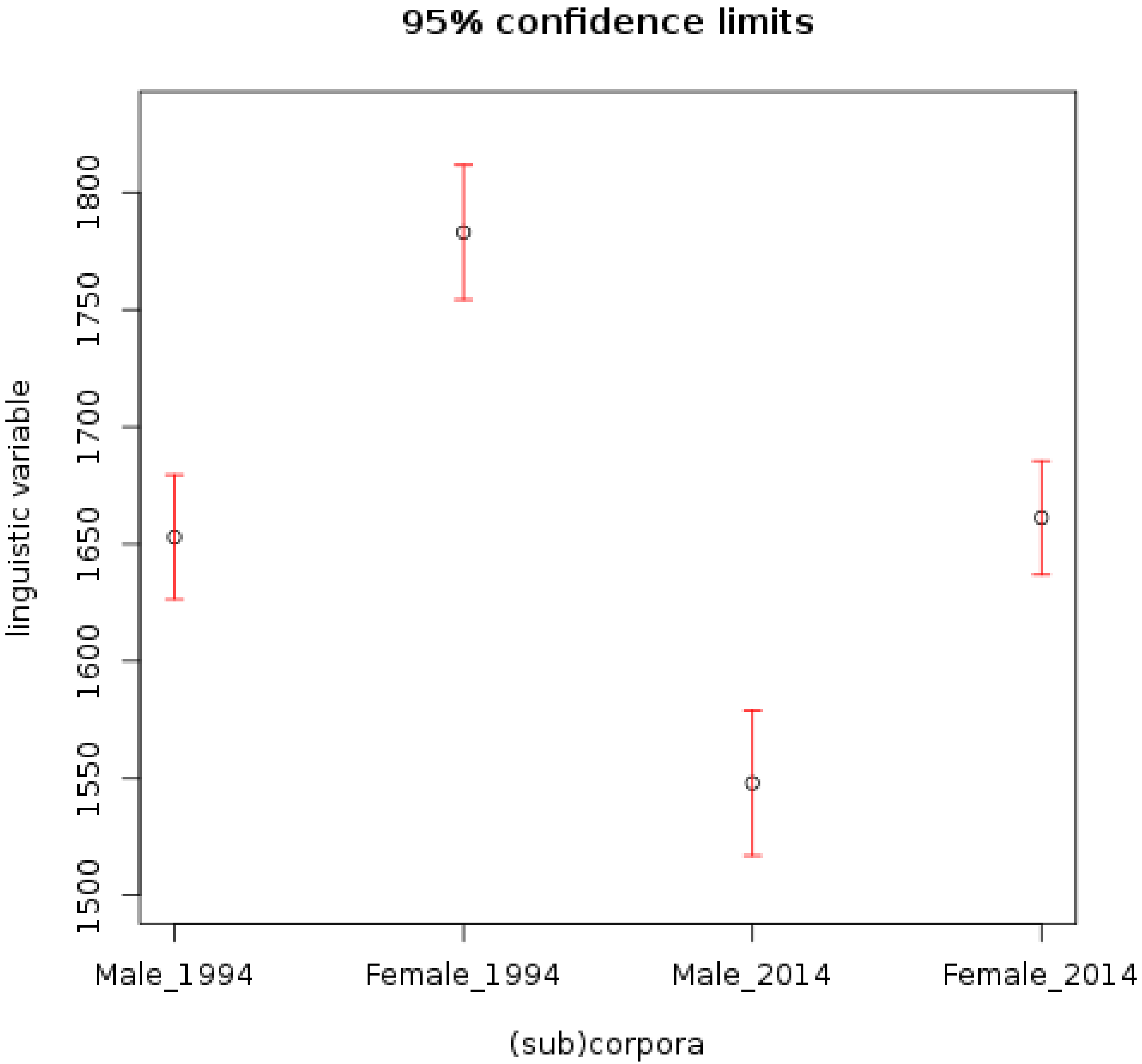
Searched KWIC for "language".

CONTEXT



Name	Tokens	MATTR$_{50}$	MTLD
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/w/index.php?title=Linguistic...	616	0.60	3.37
https://en.wikipedia.org/wiki/A_language_is_a_dialect...	577	0.81	70.96
https://en.wikipedia.org/wiki/Accent_(sociolinguistics)	2,522	0.82	97.41
https://en.wikipedia.org/wiki/Ambiguity	3,779	0.78	74.42
https://en.wikipedia.org/wiki/Applied_linguistics	922	0.73	44.21
https://en.wikipedia.org/wiki/Applied_science	931	0.79	72.38

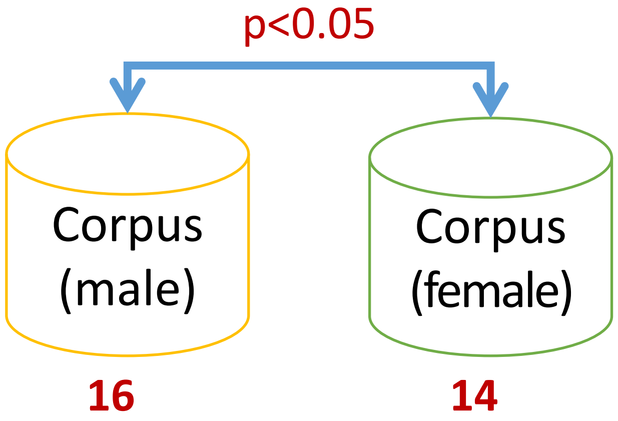
Changed value to Overview.


95% confidence limits
linguistic variable
1500
1550
1600
1650
1700
1750
1800
Male_1994
Female_1994
Male_2014
Female_2014
(sub)corpora

T-test

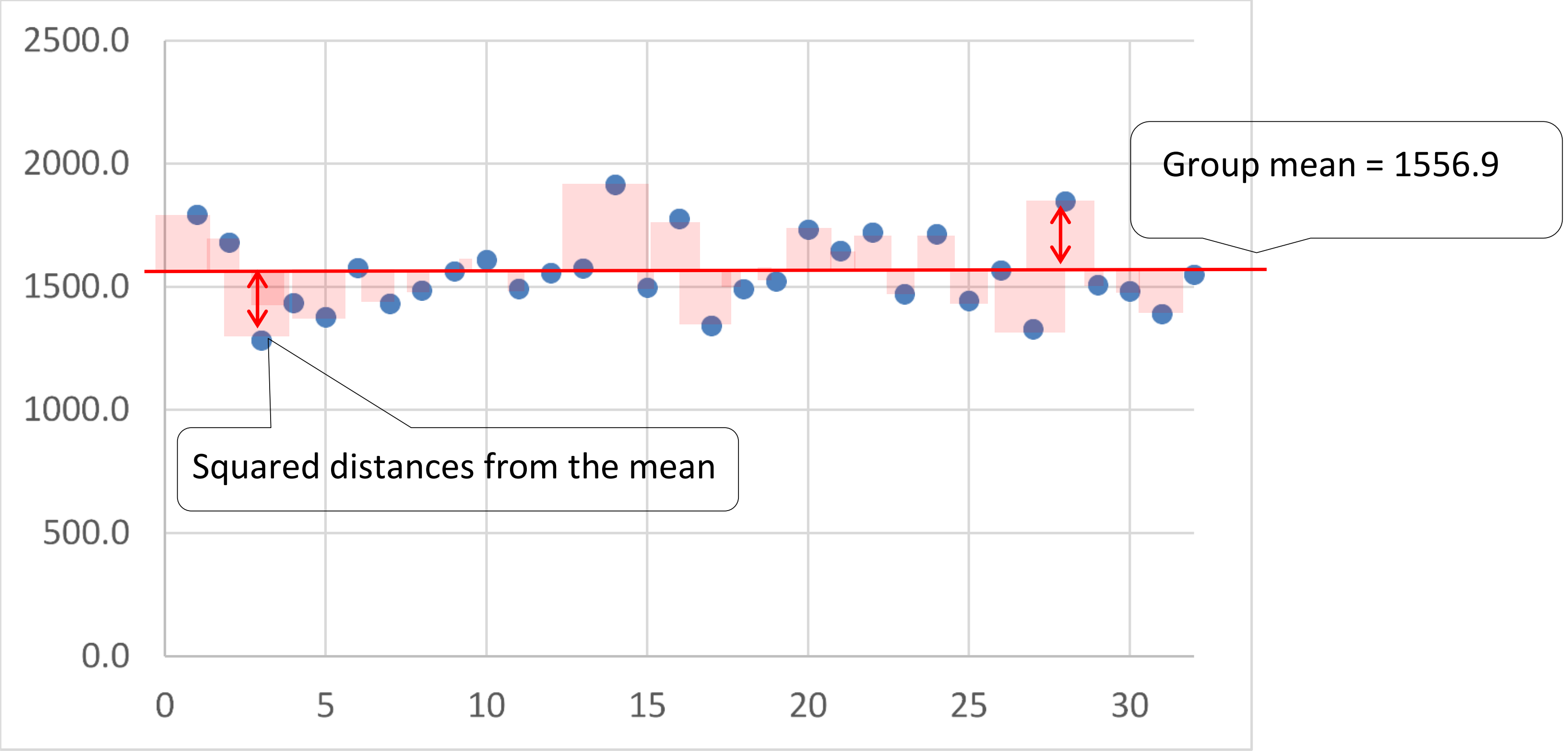
$$\text{Welch's independent sample t-test} = \frac{\text{Mean of group 1} - \text{Mean of group 2}}{\sqrt{\frac{\text{Variance of group1}}{\text{Number of cases in group 1}} + \frac{\text{Variance of group2}}{\text{Number of cases in group2}}}}$$

$$\text{Variance} = \frac{\text{sum of squared distances from the mean}}{\text{degrees of freedom}}$$



Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

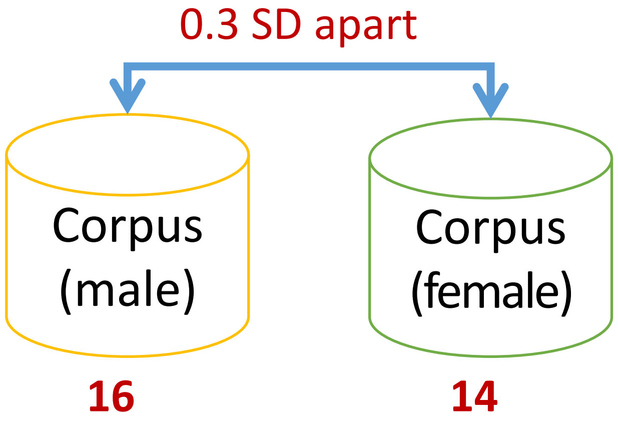
T-test (cont.)



Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Effect size: Cohen's *d*

$$\text{Cohen's } d = \frac{\text{Mean of group 1} - \text{Mean of group 2}}{\text{pooled } SD}$$



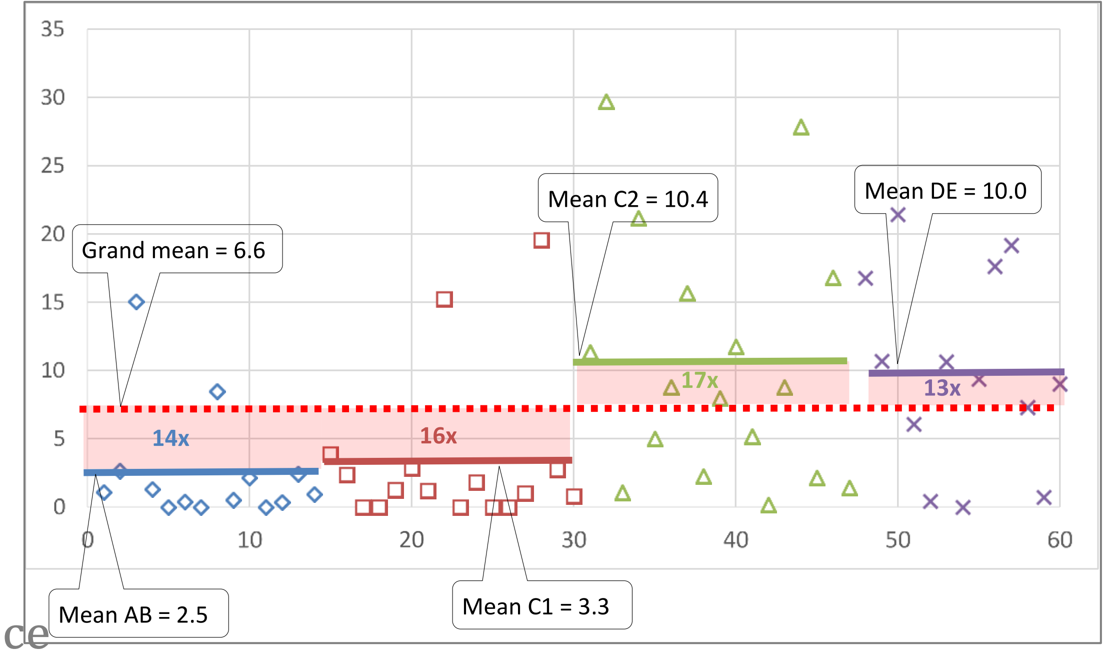
$$\text{pooled } SD = \sqrt{\frac{SD1^2 \times (\text{cases in group1} - 1) + SD2^2 \times (\text{cases in group2} - 1)}{\text{all cases} - 2}}$$

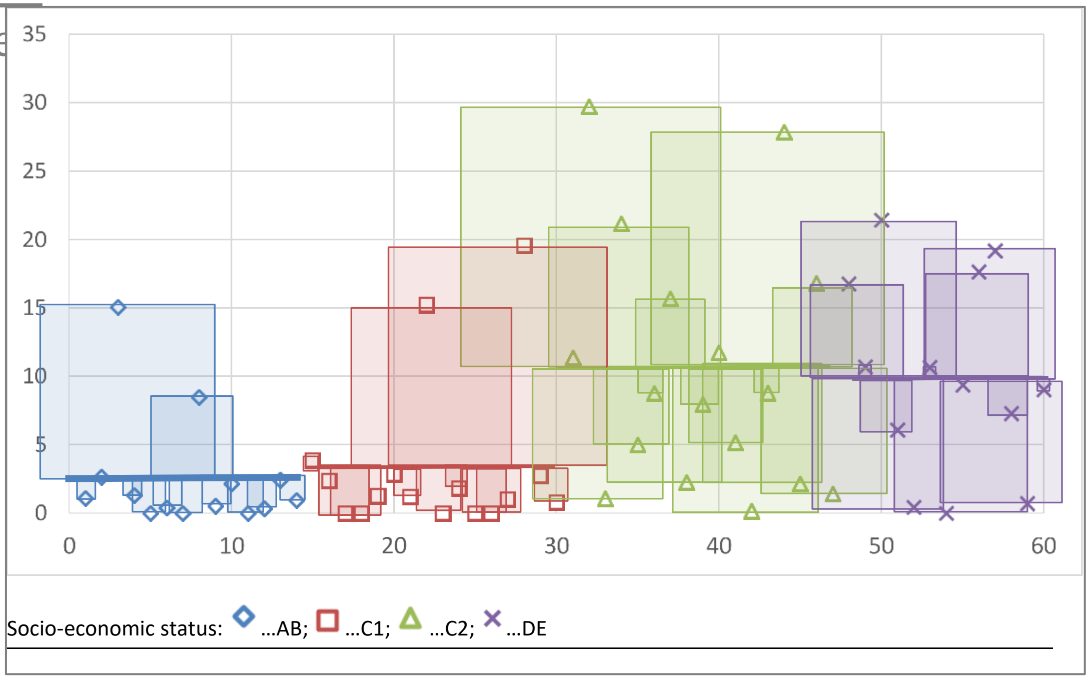
Interpretation of *d*: 0.3 small, 0.5 medium, 0.8 large effect

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

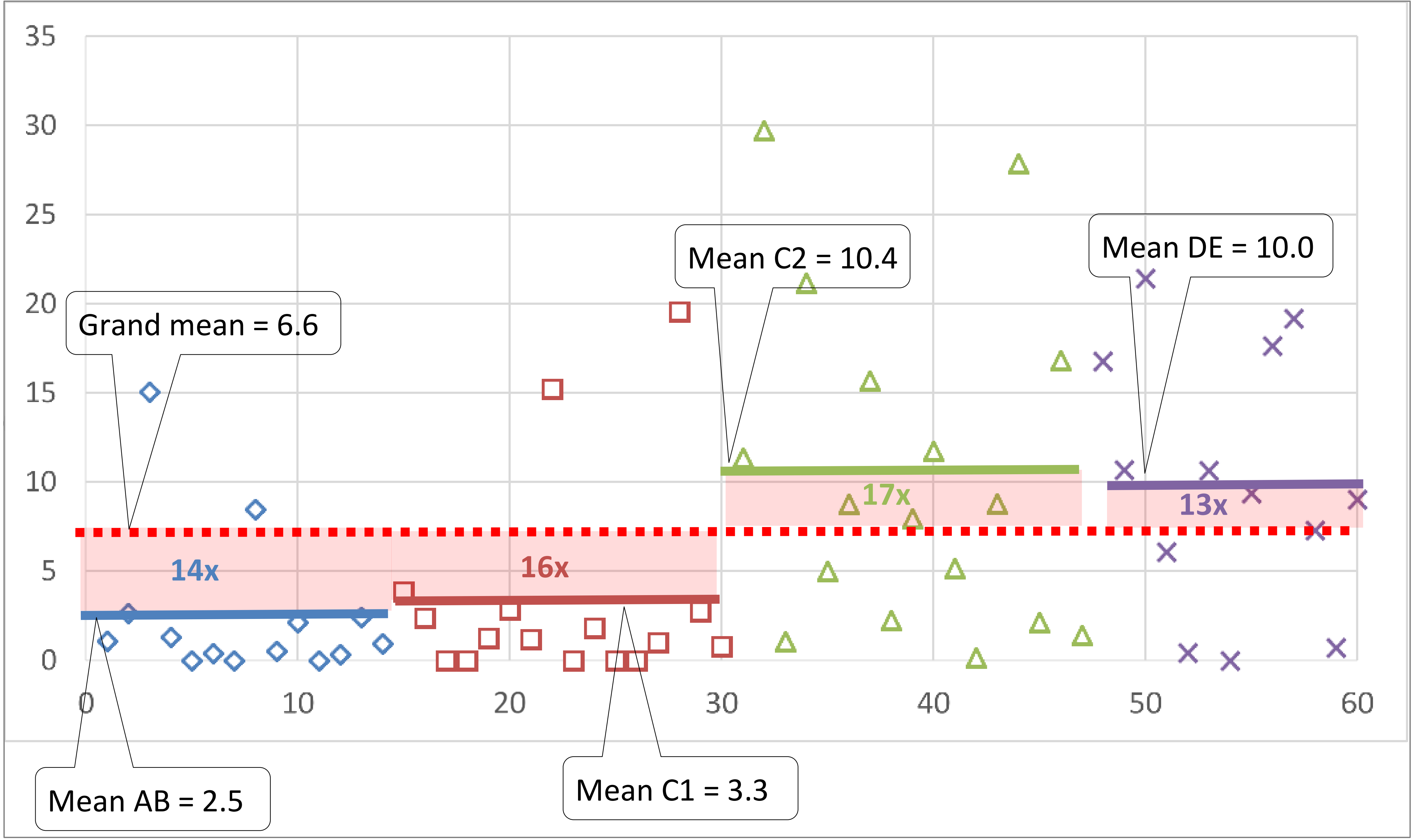
One-way ANOVA

$$\text{One-way ANOVA (F)} = \frac{\text{Between group variance}}{\text{Within group variance}}$$





Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.



Grand mean = 6.6

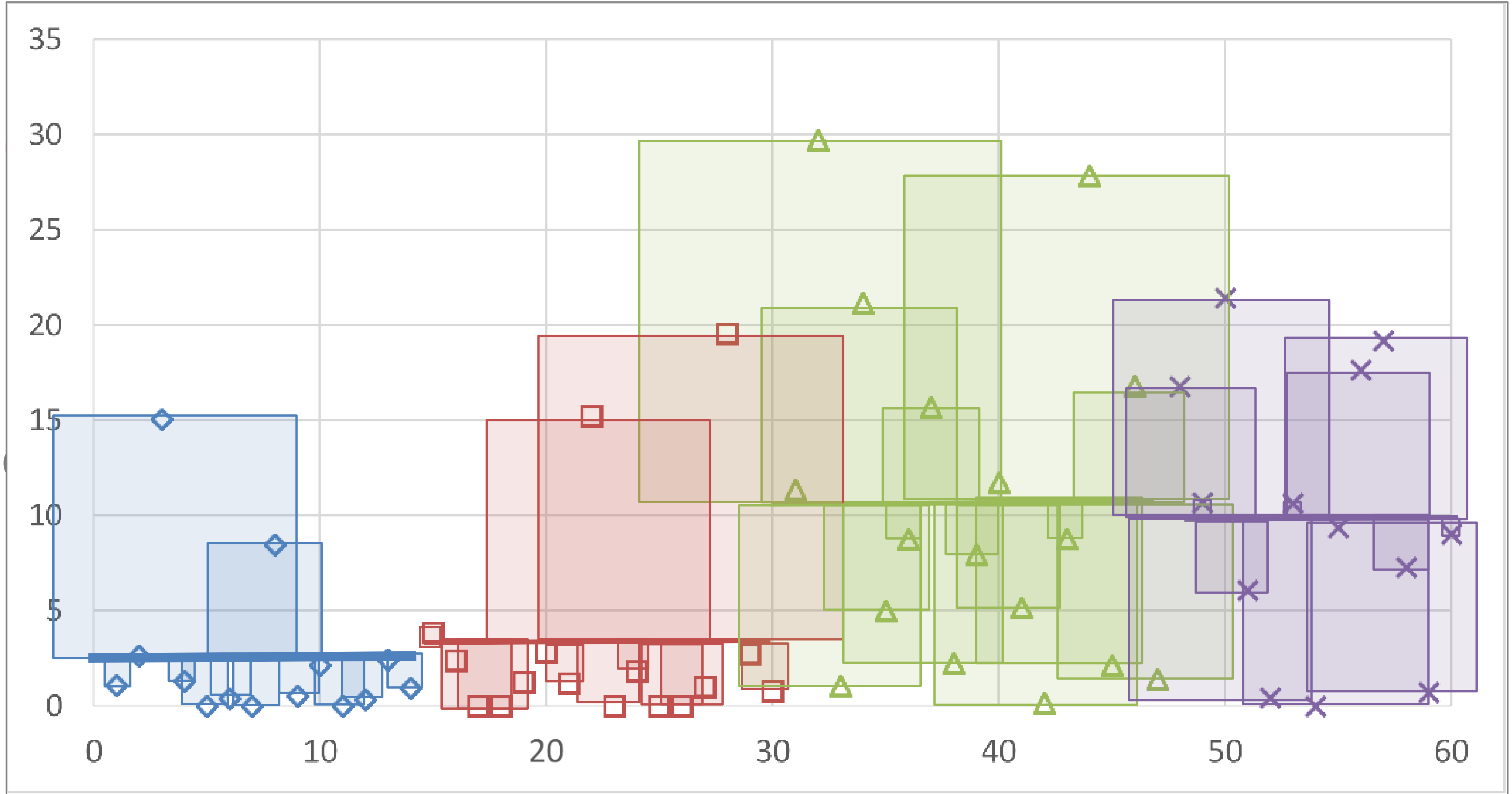
Mean C2 = 10.4

Mean DE = 10.0

Mean AB = 2.5

Mean C1 = 3.3

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.



Socio-economic status: ◇ ...AB; □ ...C1; △ ...C2; × ...DE

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Tests: overview

Test	**T-test**	**ANOVA**	**Mann-Whitney U**	**Kruskal-Wallis**
No. of groups compared	2	2+	2	2+
Assumes underlying normal distribution of the linguistic variable in the population	YES	YES	NO	NO
Assumes independence of texts/speakers	YES	YES	YES	YES
Allows testing interaction between different explanatory variables (e.g. register and author's gender)	NO	YES	NO	NO

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.