

LARGE LINGUISTIC NETWORKS

A CASE STUDY ON LONE-WOLF TERRORISM IN THE BRITISH PRESS



Hanna Schmück
Lancaster University



WHO AM I?



Hanna Schmück

Corpus Linguist

Associate Lecturer, Research Assistant, and
Doctoral Candidate at Lancaster University

Areas of interest:

collocation analysis, computational
approaches, psycholinguistics, explainable
methodologies

PART I

+

•

○



WHAT IS CORPUS LINGUISTICS?



WHAT ARE COLLOCATIONS?



WHAT CAN WE EXPLORE USING
COLLOCATION NETWORKS?

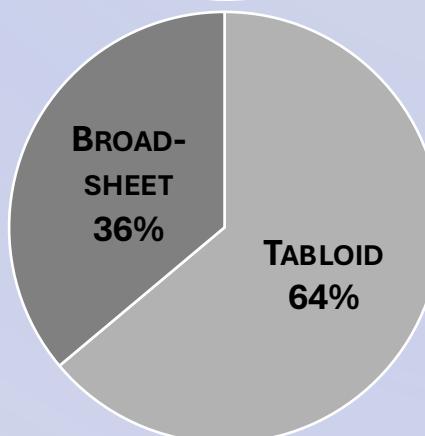
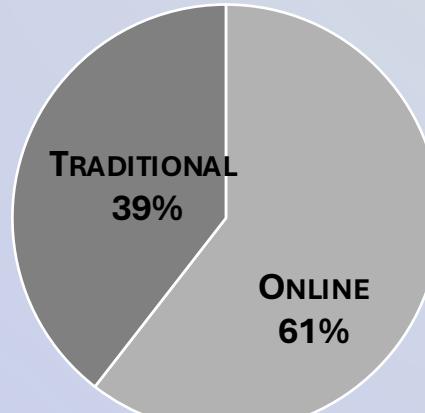
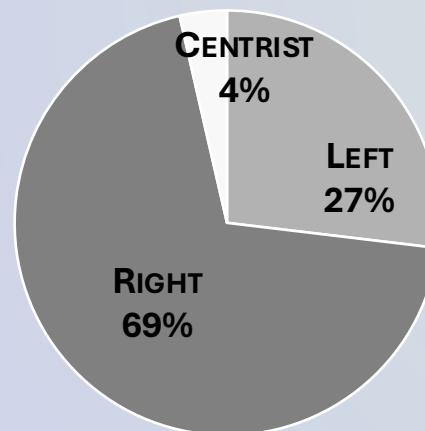
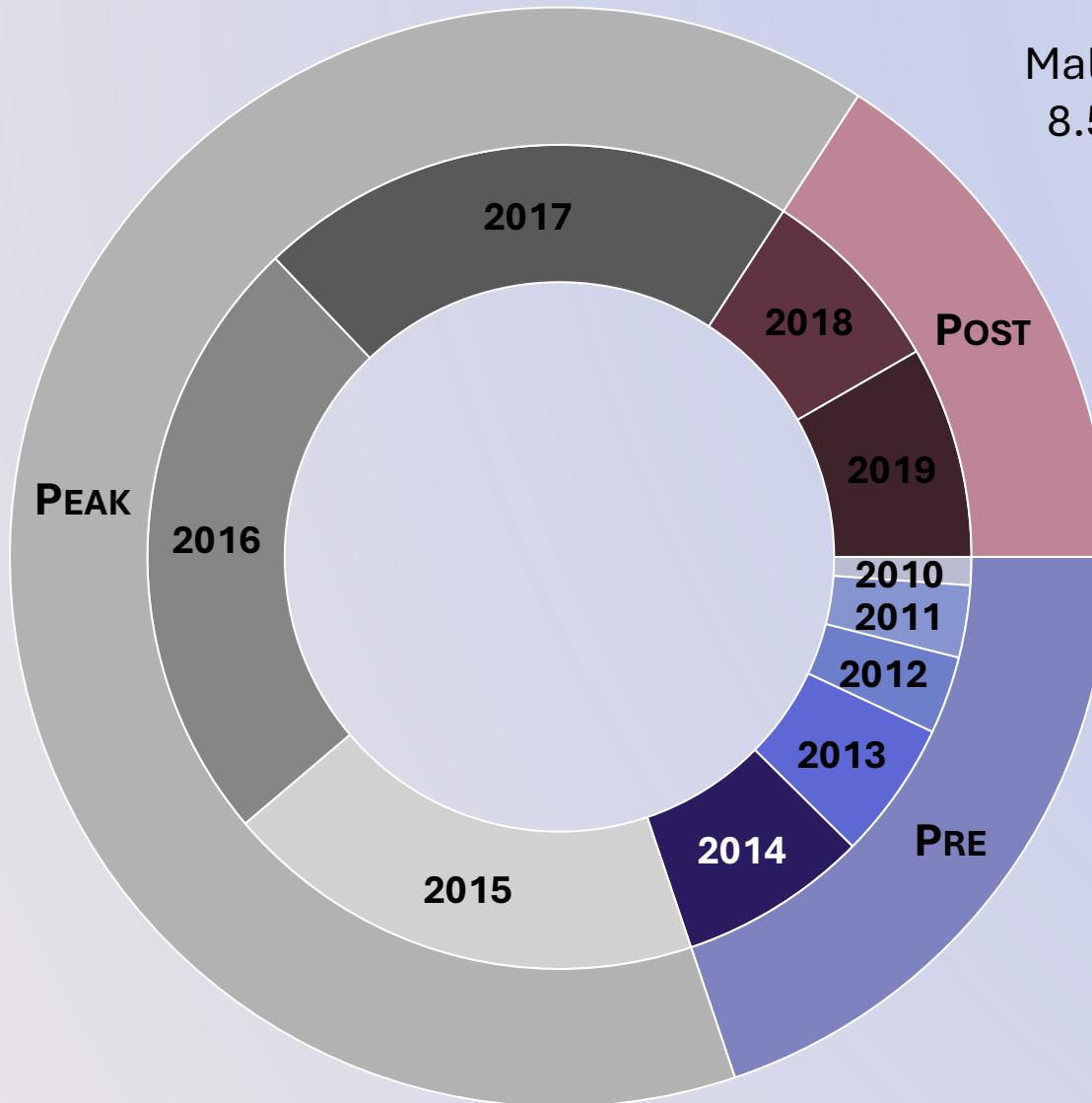
WHAT IS CORPUS LINGUISTICS? AND WHAT IS A CORPUS?

A CORPUS IS...

- A LARGE BODY OF TEXT
- AIMS TO BE REPRESENTATIVE OF LANGUAGE (OR A GENRE OF LANGUAGE)
- MACHINE READABLE (MCENERY & HARDIE, 2011)

OFTEN ANNOTATED WITH ADDITIONAL LINGUISTIC INFORMATION – E.G. GRAMMATICAL CODES,
LEARNER ERRORS ETC.

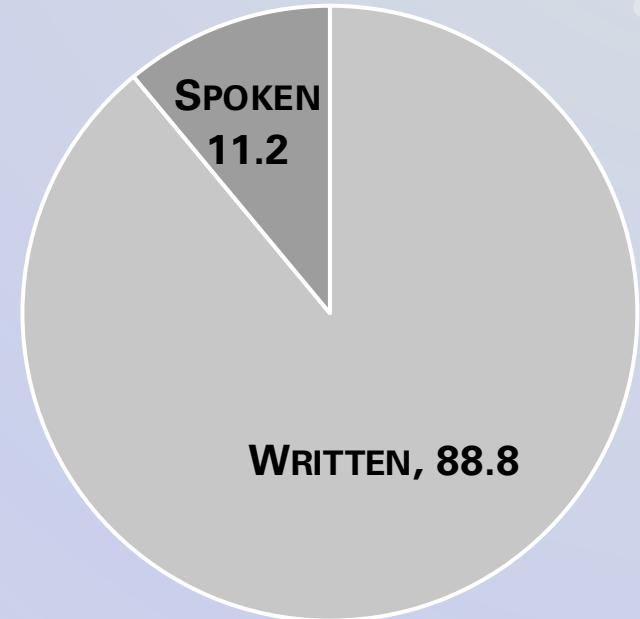
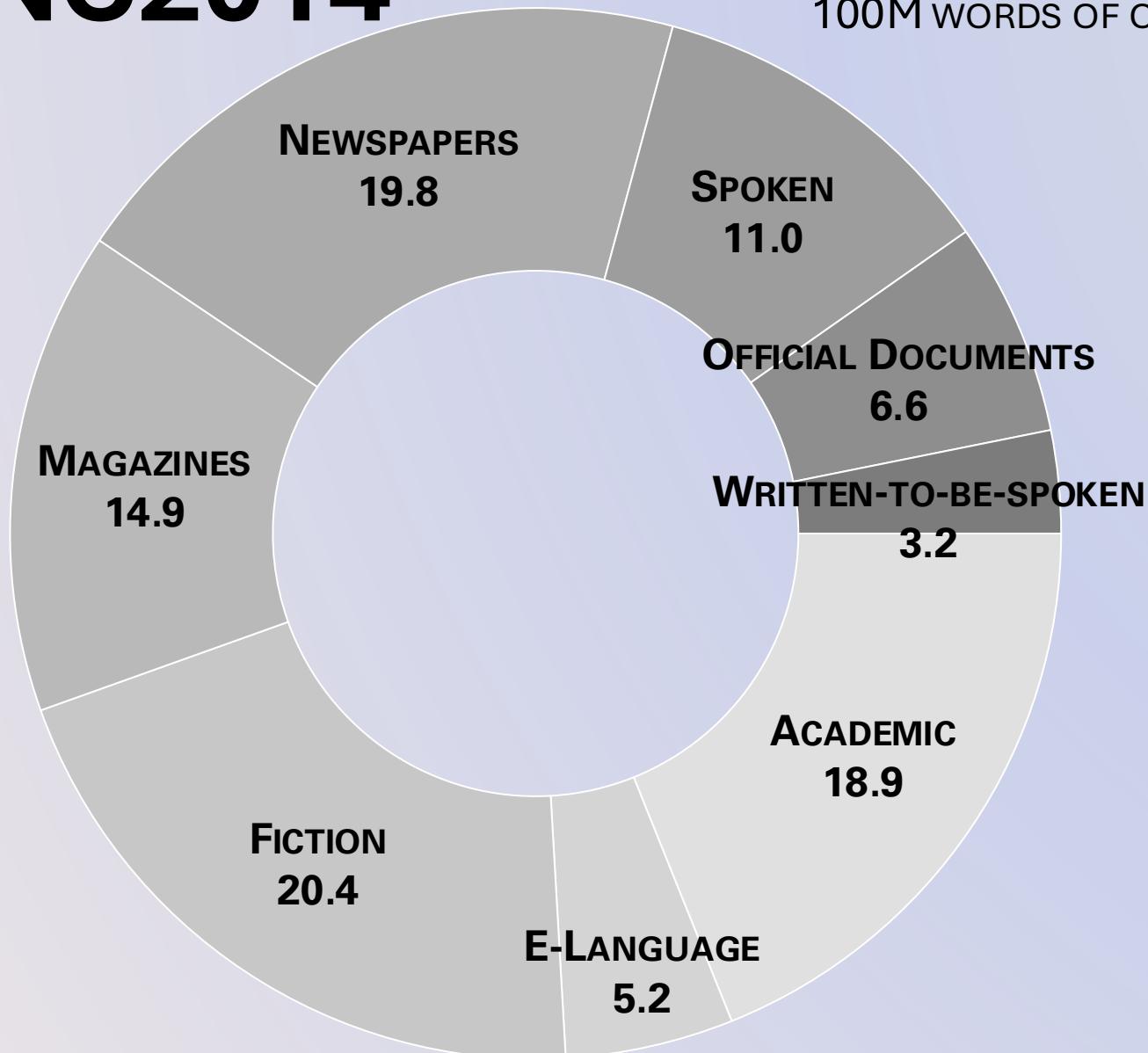
THE LONE WOLF CORPUS



THE BNC2014

100M WORDS OF CONTEMPORARY BRITISH ENGLISH

(BREZINA ET AL., 2021; LOVE ET AL., 2017)



WHAT IS A COLLOCATION?

- commonly co-occurring group or set of words (BARNBROOK ET AL., 2013, P. 3; STULPINAITĖ ET AL., 2016, P. 31)

WHAT DO THEY REPRESENT?

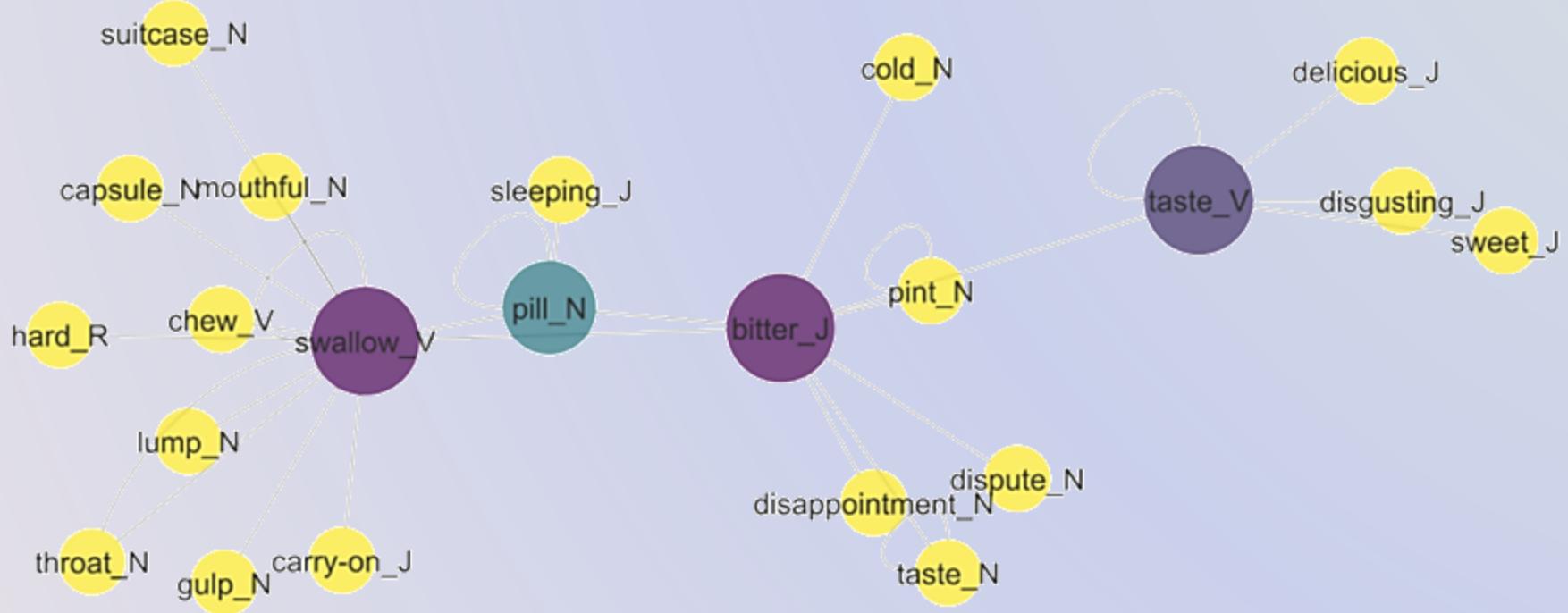
- language learning context: high language proficiency and fluency
- Mental Lexicon: conventionalised and entrenched form-meaning mappings (CROFT & CRUSE, 2004, P. 292; SIMPSON-VLACH & ELLIS, 2010, P. 488)
- contextual embeddings – can be used to track changes over time
(SEE MCGILLIVRAY & TÓTH (2020) FOR AN EXAMPLE)

WHAT IS A COLLOCATION?

- BREAK THE CAR|WINDOW|ICE|PEACE
- IT'S NO USE CRYING OVER SPILT MILK
- MONKEY'S BIRTHDAY
- MOTHER EARTH
- FATHER CHRISTMAS

BUT ALSO GRAMMATICAL COLLOCATIONS: GOING TO, FEELING LIKE ETC.

EXAMPLE



idiom *[leave a] bitter taste and a bitter pill to swallow*

WHY USE COLLOCATION NETWORKS?

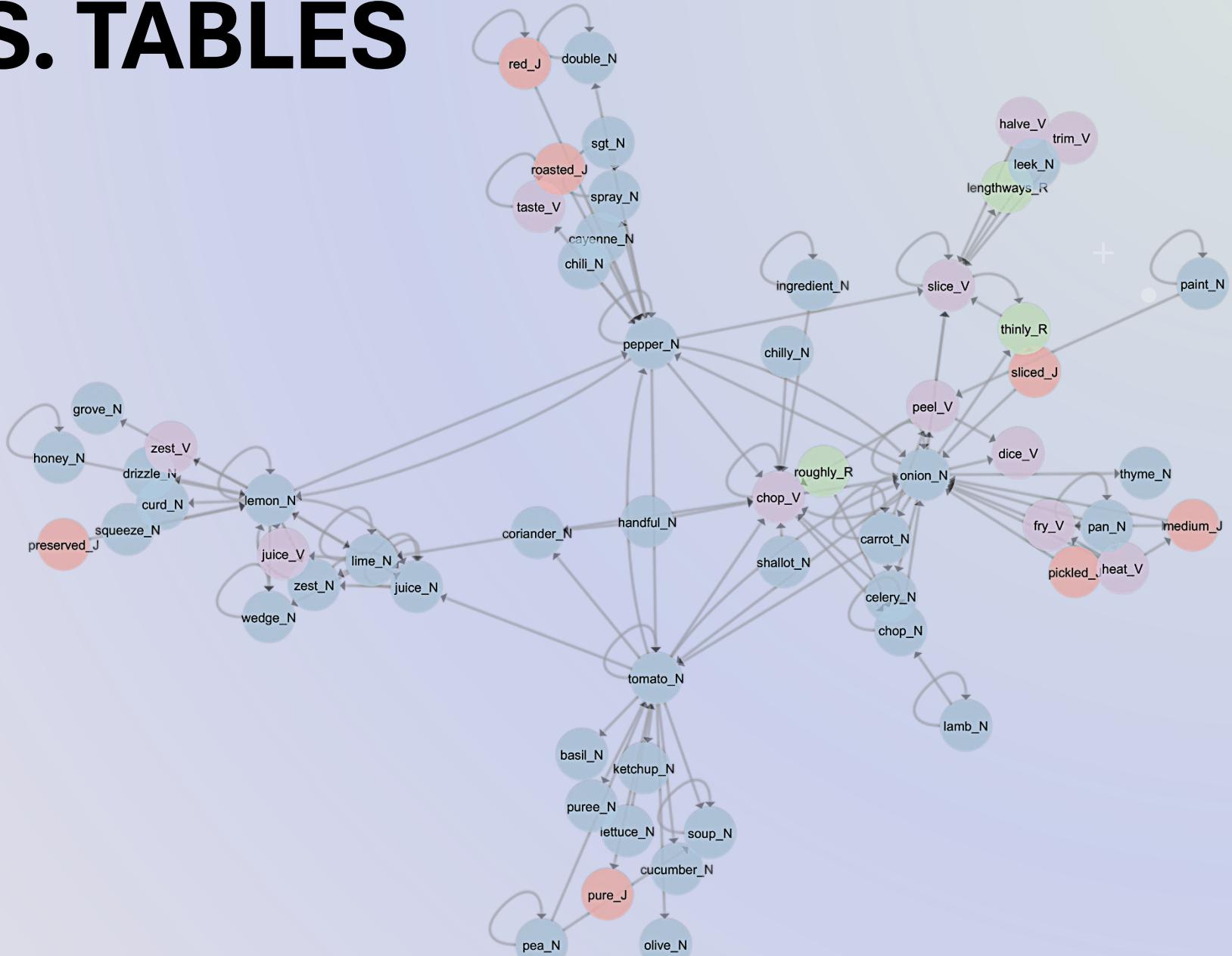
Visualisation of

- **discourse relationships, aboutness** of a text or discourse
- **semantic relations** (Pecina, 2010; Xiao & McEnery, 2006; Brezina, 2016; Baker, 2016; Brezina et al., 2015)
- **lexicogrammatical features** (McEnery & Brezina, 2019)

NETWORKS VS. TABLES

node	collocate	logDice	node	collocate	logDice	node	collocate	logDice
thinly_R	sliced_J	10.38	lime_N	lime_N	8.10	chop_V	onion_N	7.46
lemon_N	juice_N	10.14	heat_V	fry_V	8.07	chop_V	chop_V	7.44
lemon_N	zest_N	10.07	lemon_N	zest_V	8.05	tomato_N	cucumber_N	7.43
thinly_R	slice_V	10.07	carrot_N	peel_V	8.04	onion_N	carrot_N	7.42
onion_N	chop_V	9.26	lemon_N	lemon_N	7.99	pickled_J	onion_N	7.41
heat_V	pan_N	9.25	taste_V	taste_V	7.95	chilly_N	chop_V	7.41
chop_N	chop_N	9.17	chop_V	chop_N	7.94	pepper_N	taste_V	7.40
pea_N	pea_N	9.10	cayenne_N	pepper_N	7.94	celery_N	chop_V	7.38
roughly_R	chop_V	9.02	juice_N	lime_N	7.93	trim_V	slice_V	7.36
lime_N	juice_N	8.95	onion_N	pepper_N	7.92	pepper_N	slice_V	7.32
onion_N	slice_V	8.90	juice_N	zest_N	7.91	onion_N	celery_N	7.31
carrot_N	celery_N	8.88	lemon_N	drizzle_N	7.88	roasted_J	pepper_N	7.31
shallot_N	chop_V	8.84	handful_N	coriander_N	7.87	chop_N	onion_N	7.30
wedge_N	wedge_N	8.83	onion_N	tomato_N	7.86	sgt_N	pepper_N	7.27
lime_N	juice_V	8.79	handful_N	chop_V	7.83	lemon_N	grove_N	7.26
lemon_N	juice_V	8.71	onion_N	peel_V	7.83	onion_N	thinly_R	7.26
juice_N	lemon_N	8.70	tomato_N	pepper_N	7.80	tomato_N	chop_V	7.22
lamb_N	chop_N	8.70	peel_V	chop_V	7.79	tomato_N	pure_J	7.22
peel_V	slice_V	8.62	carrot_N	onion_N	7.79	onion_N	thyme_N	7.20
soup_N	soup_N	8.60	lemon_N	lime_N	7.78	preserved_J	lemon_N	7.19
tomato_N	ketchup_N	8.59	pepper_N	onion_N	7.76	honey_N	lemon_N	7.18
honey_N	honey_N	8.51	pepper_N	tomato_N	7.76	pan_N	pan_N	7.16
carrot_N	carrot_N	8.48	zest_N	juice_N	7.75	slice_V	slice_V	7.16
tomato_N	tomato_N	8.45	leek_N	slice_V	7.73	ingredient_N	chop_V	7.15
slice_V	thinly_R	8.37	tomato_N	puree_N	7.71	lemon_N	pepper_N	7.10
coriander_N	lime_N	8.34	paint_N	paint_N	7.70	roughly_R	chop_N	7.10
slice_V	lengthways_R	8.32	peel_V	dice_V	7.69	heat_V	onion_N	7.10
spray_N	spray_N	8.30	coriander_N	chop_V	7.67	tomato_N	coriander_N	7.09
zest_N	lemon_N	8.28	sliced_J	onion_N	7.65	tomato_N	olive_N	7.07
pan_N	fry_V	8.28	tomato_N	juice_N	7.62	pea_N	soup_N	7.03
lime_N	wedge_N	8.27	squeeze_N	lemon_N	7.62	carrot_N	tomato_N	7.02
fry_V	onion_N	8.26	pan_N	onion_N	7.60	pea_N	tomato_N	7.02
lime_N	zest_N	8.24	heat_V	medium_J	7.59	double_N	double_N	7.02
lamb_N	lamb_N	8.22	lettuce_N	tomato_N	7.59	pepper_N	double_N	7.01
pan_N	medium_J	8.22	pepper_N	spray_N	7.57	pepper_N	lemon_N	7.01
tomato_N	basil_N	8.21	juice_N	juice_N	7.57	red_J	red_J	7.01
onion_N	dice_V	8.18	pepper_N	pepper_N	7.57	medium_J	onion_N	7.01
chili_N	pepper_N	8.14	lemon_N	wedge_N	7.54	paint_N	peel_V	6.99
tomato_N	onion_N	8.13	tomato_N	soup_N	7.51	red_J	pepper_N	6.98
onion_N	onion_N	8.12	pepper_N	chop_V	7.50	ingredient_N	ingredient_N	6.98
lemon_N	curd_N	8.11	halve_V	slice_V	7.47			

NETWORKS VS. TABLES



(Atomic unit: lemma_POS, AM: logDice,

Threshold: 6.73, Sentence-span, min

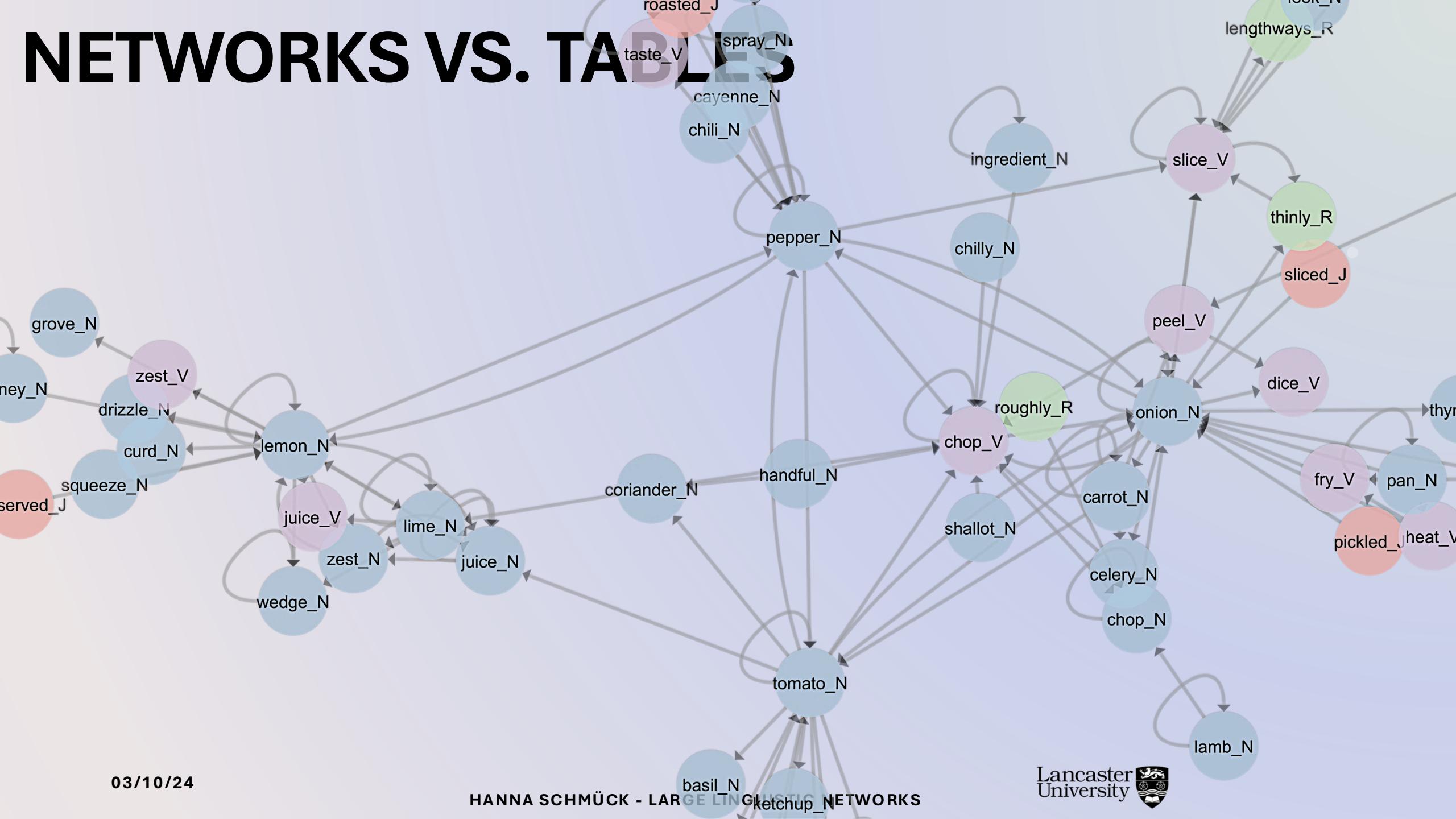
collocation frequency: 10, min collocate

frequency: 10)

(2-dimensional, Edge length: AM, Colour-coding: POS, Layout type: Edge-weighted spring embedded) – Visualisation Software:

Cytoscape (Shannon, 2003)

NETWORKS VS. TABLES



LARGE LINGUISTIC NETWORKS

General idea: bridging the gap between **corpus linguistics** and **graph theory**

Graph theory (Gross et al., 2019) is a branch of mathematics focusing on nodes (points) connected by edges (lines) and their properties

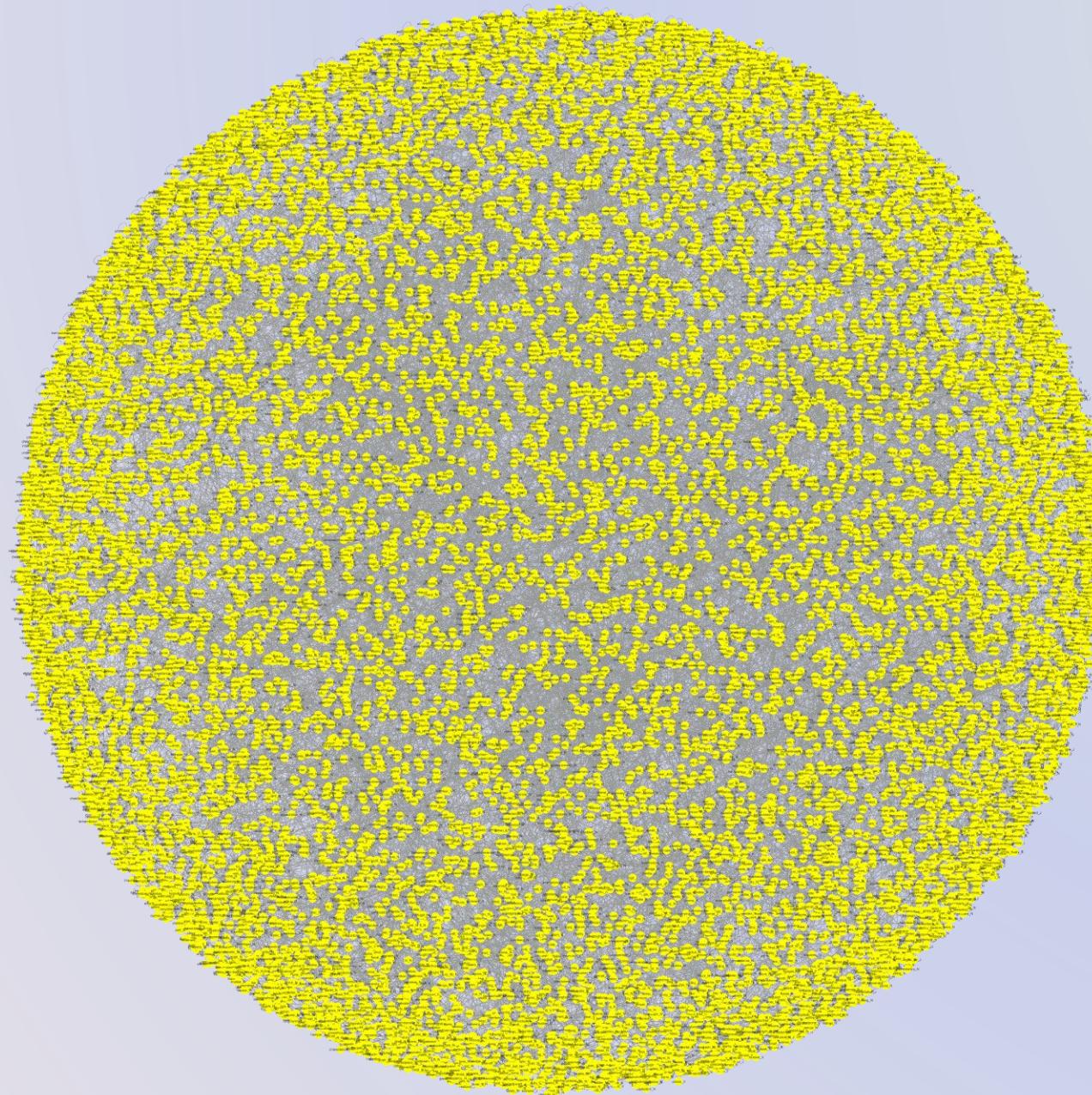
LLNs can be generated on different linguistic data, here we focus on corpus-wide **collocations**

LLN

BNC-2014 – LOGDICE > 6.73

LARGEST CONNECTED COMPONENT

NODE COLOUR = POS



Atomic unit: lemma_POS, AM:
logDice, Threshold: 6.73,
Sentence-span, min collocation
frequency: 10, min collocate
frequency: 10

2-dimensional, Edge length: AM,
Colour-coding: POS, Layout type:
Edge-weighted spring embedded

LLN

03/10/24

HANNA SCHMUCK - LARGE LINGUISTIC NETWORKS

3 OCTOBER 2024

EXPLORING THE NEXUS BETWEEN THE LONE-WOLF TERRORIST, AL- QAEDA, AND ISIS IN THE BRITISH PRESS

PART II

HANNA SCHMÜCK - LARGE LINGUISTIC NETWORKS





REPRESENTATION OF THE CORPUS-WIDE COLLOCATION NETWORKS BASED ON THE UNCLUSTERED PRE-PEAK CORPUS.

THE PIPELINE

<https://osf.io/mw4jt/>
<https://doi.org/10.17605/OSF.IO/MW4J>

T

A pack of lone wolves

lone_wolf_bigrams_script.ipynb



Lone Wolf Networks

This ipynb documents the procedure used to generate collocation networks for the lone wolf project. In line with existing key word analyses, MI scores and log likelihood of bigrams have been extracted from a diachronic corpus on lone wolf discourses in the British Press. Years have been grouped into pre-peak, peak, and post-peak (see classification below).

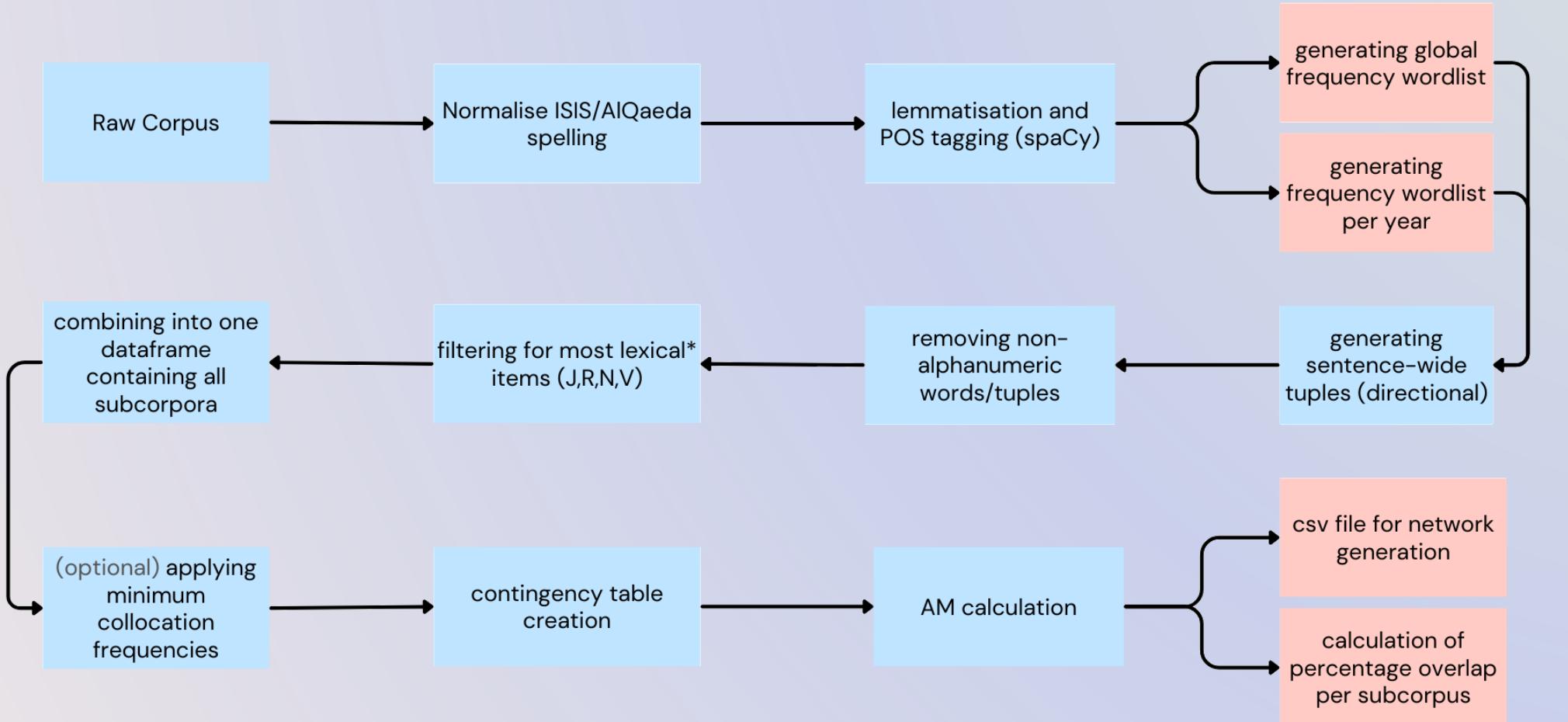
Key information regarding the bigram generation:

- bigrams cannot span multiple sentences
- Filters and thresholds are fully customisable, to keep collocation networks at a manageable size and to mirror ANON's previous work a minimum LL score of 50 had been implemented

In [1]:

```
from matplotlib.ticker import FuncFormatter
import os
import re
import random
import pickle
import pandas as pd
import numpy as np
from collections import Counter
```

THE PIPELINE



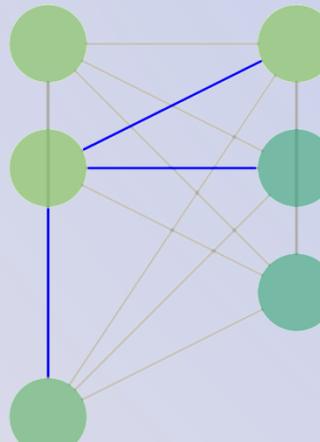
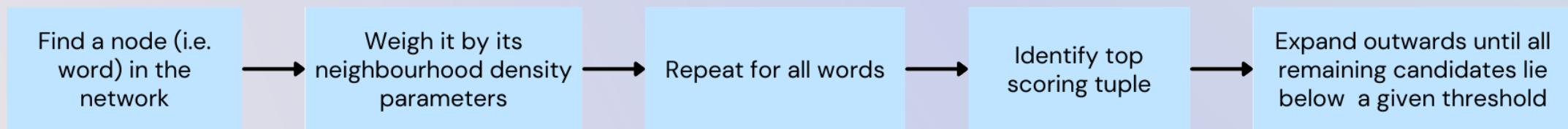
METHODOLOGY

LLN Collocation analysis

- Selection of relevant layer(s) of analysis
 - here: lemma_POS, AM: logLikelihood > 50, sentence-span
- Selection of relevant properties
 - here: density, centrality, cluster memberships
- Selection of clustering algorithm
 - here: MCODE clustering (Bader & Hogue, 2003)
- Selection of appropriate layout
 - here: edge-weighted spring directed layout (Kamada & Kawai, 1989) based on AM strength

CLUSTERING

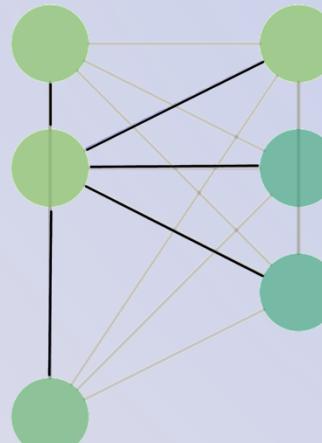
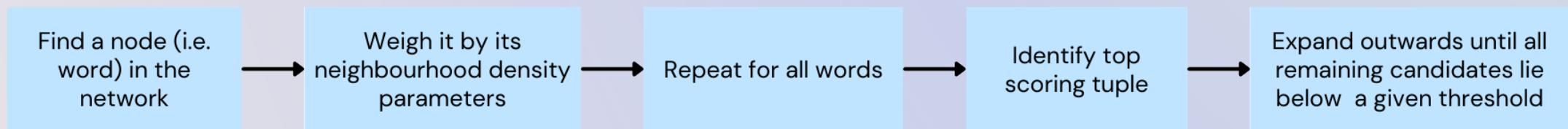
- Here MCODE clustering (Bader & Hogue, 2003)
- In simple terms, the clustering algorithm follows these steps:



3 actual connections

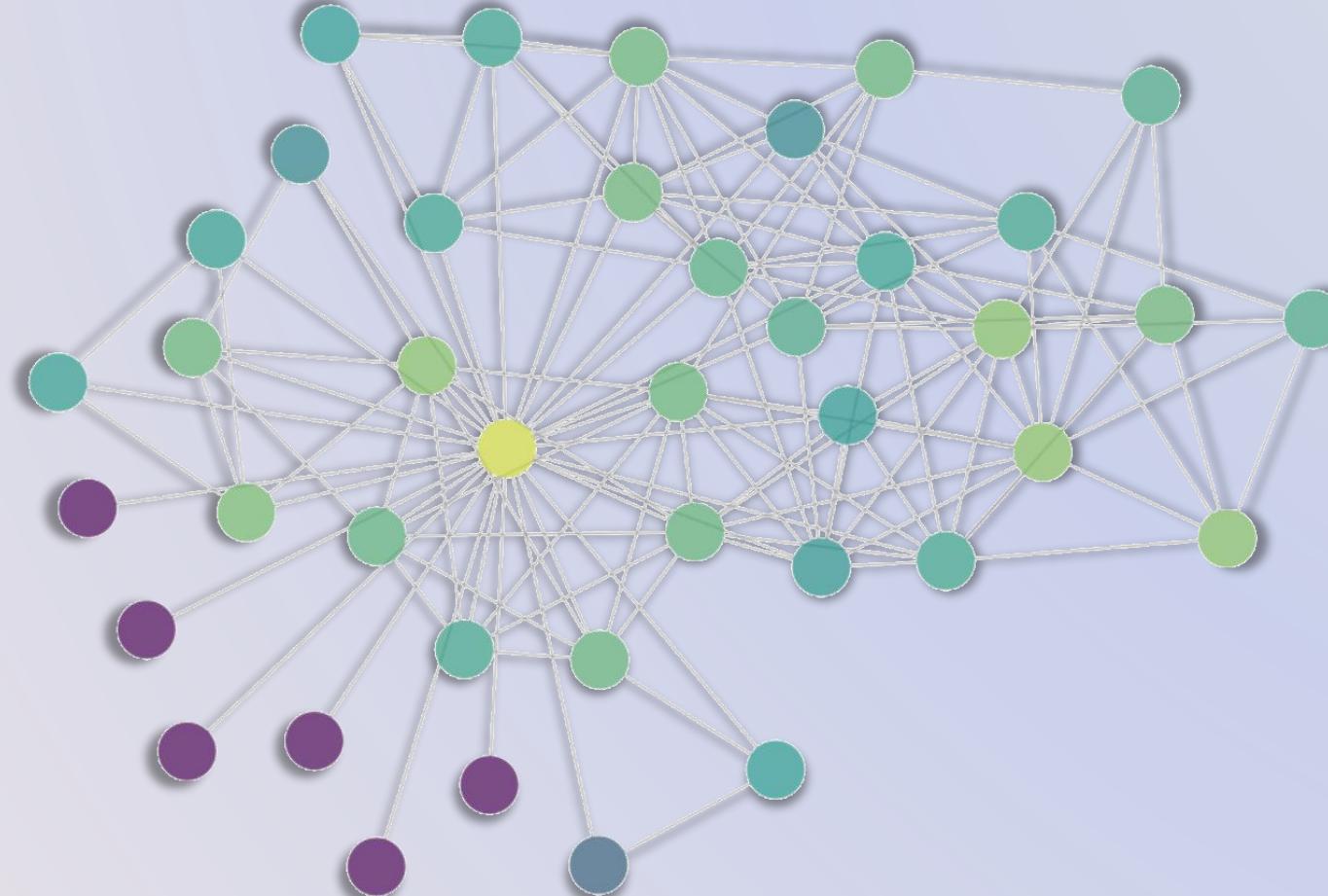
CLUSTERING

- Here MCODE clustering (Bader & Hogue, 2003)
- In simple terms, the clustering algorithm follows these steps:

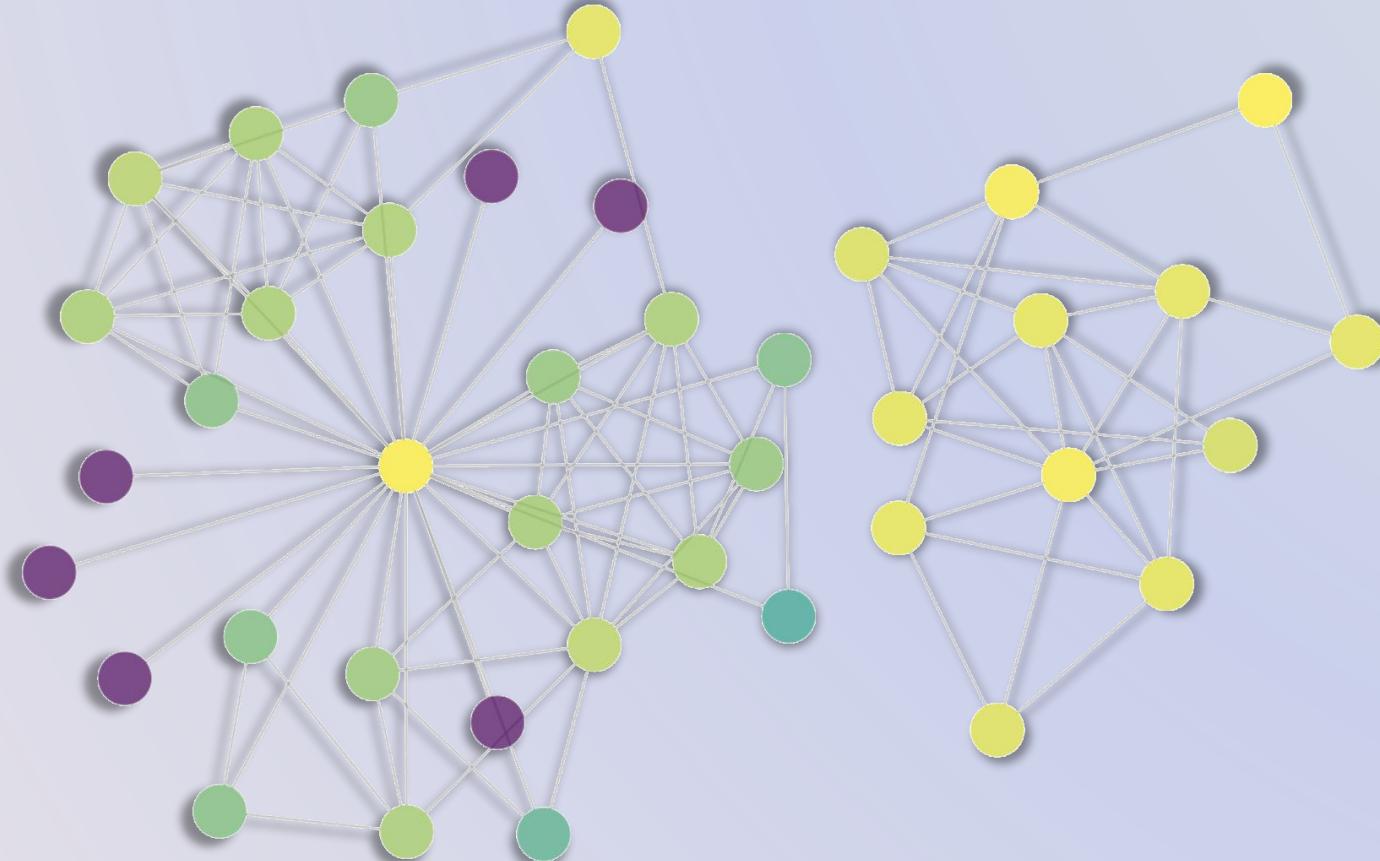


5 possible connections

CLUSTERING



CLUSTERING

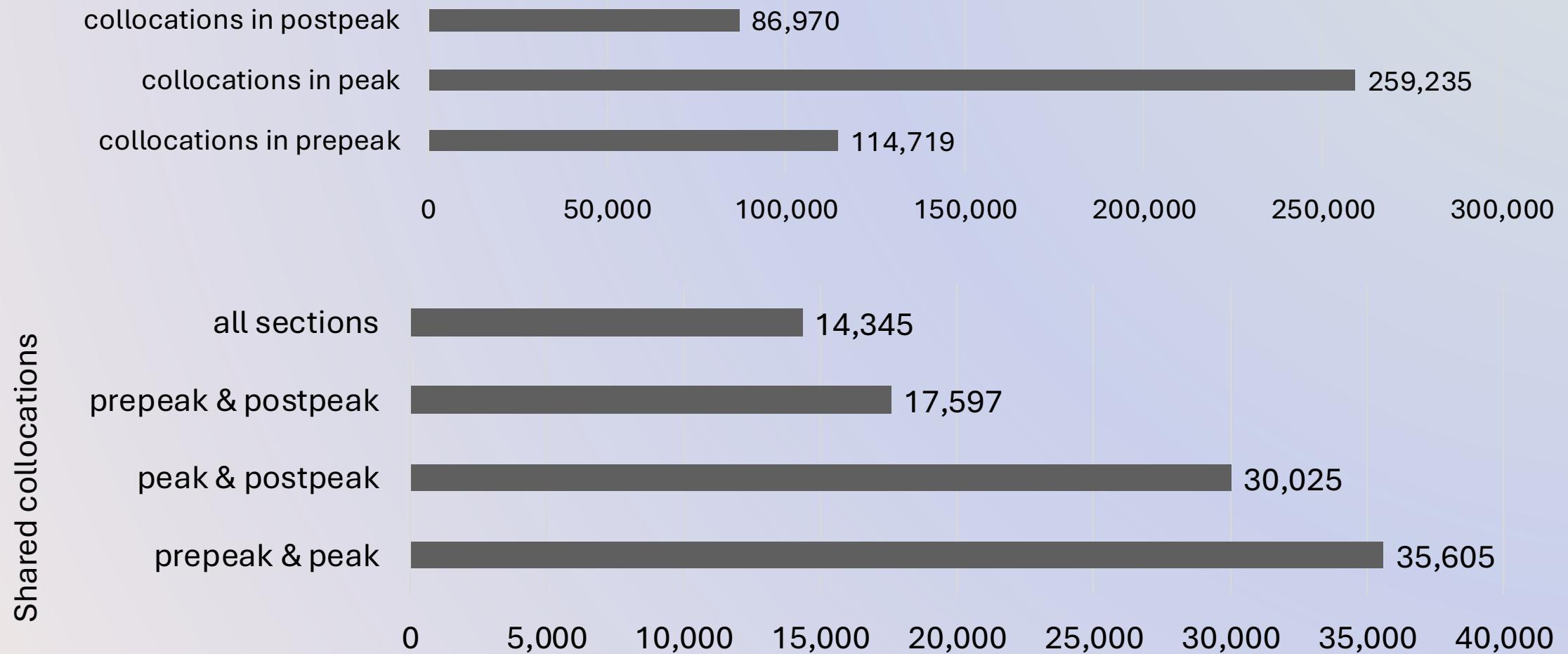


LWC CLUSTER RESULTS

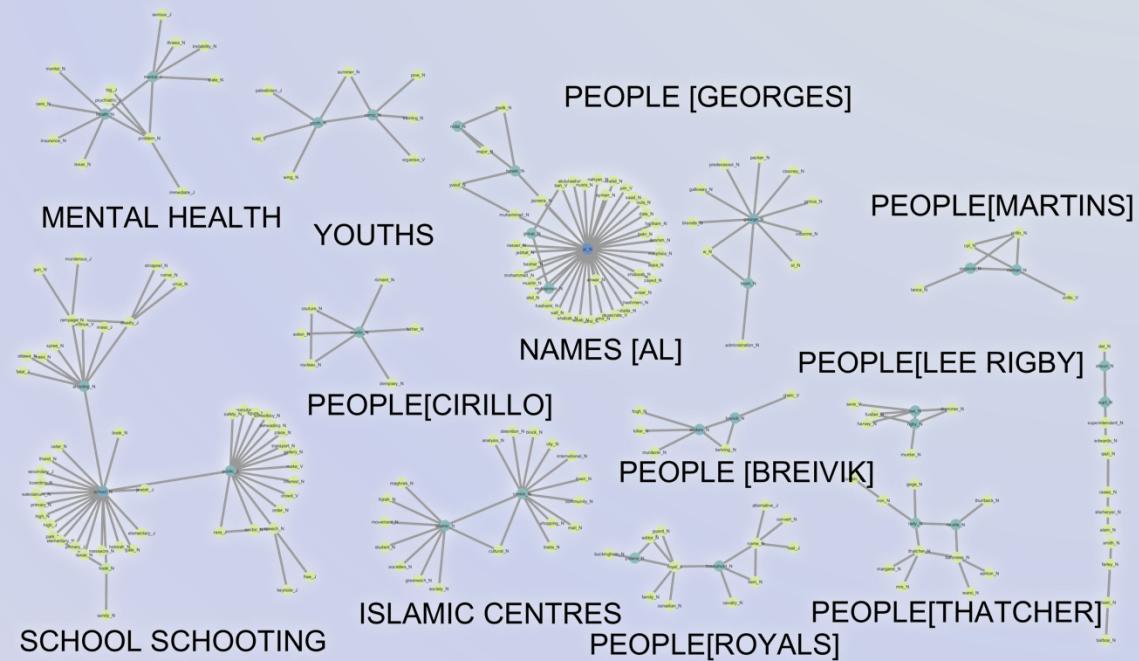
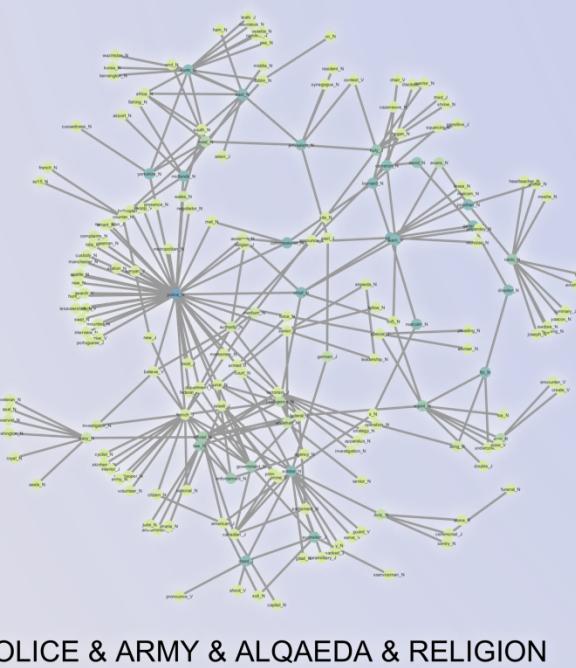
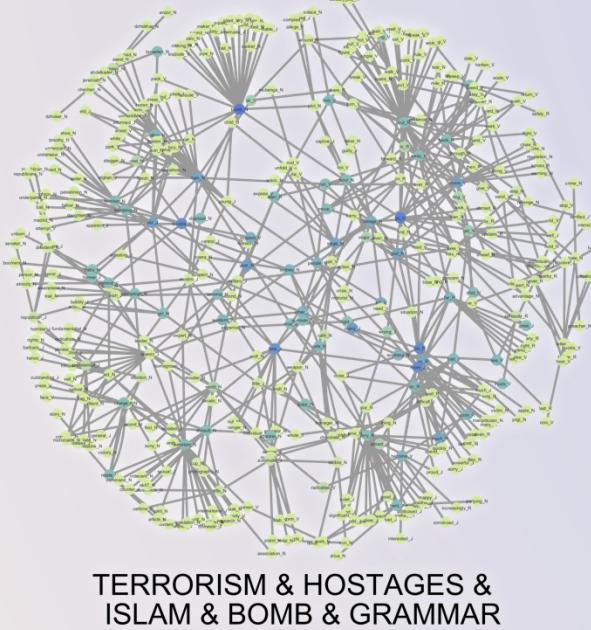
LWC NETWORK PROPERTIES

	PREPEAK	PEAK	POSTPEAK
Number of nodes	6,538	13,264	5,387
Number of edges	7,184	20,695	5,741
Avg. number of neighbours	2.2	3.1	2.1
Network diameter	27	22	29
Characteristic path length	9.4	6.8	9.5
Clustering coefficient	0.006	0.01	0.007
Connected components	1,068	1,473	866

SUBCORPUS SIMILARITY



PREPEAK

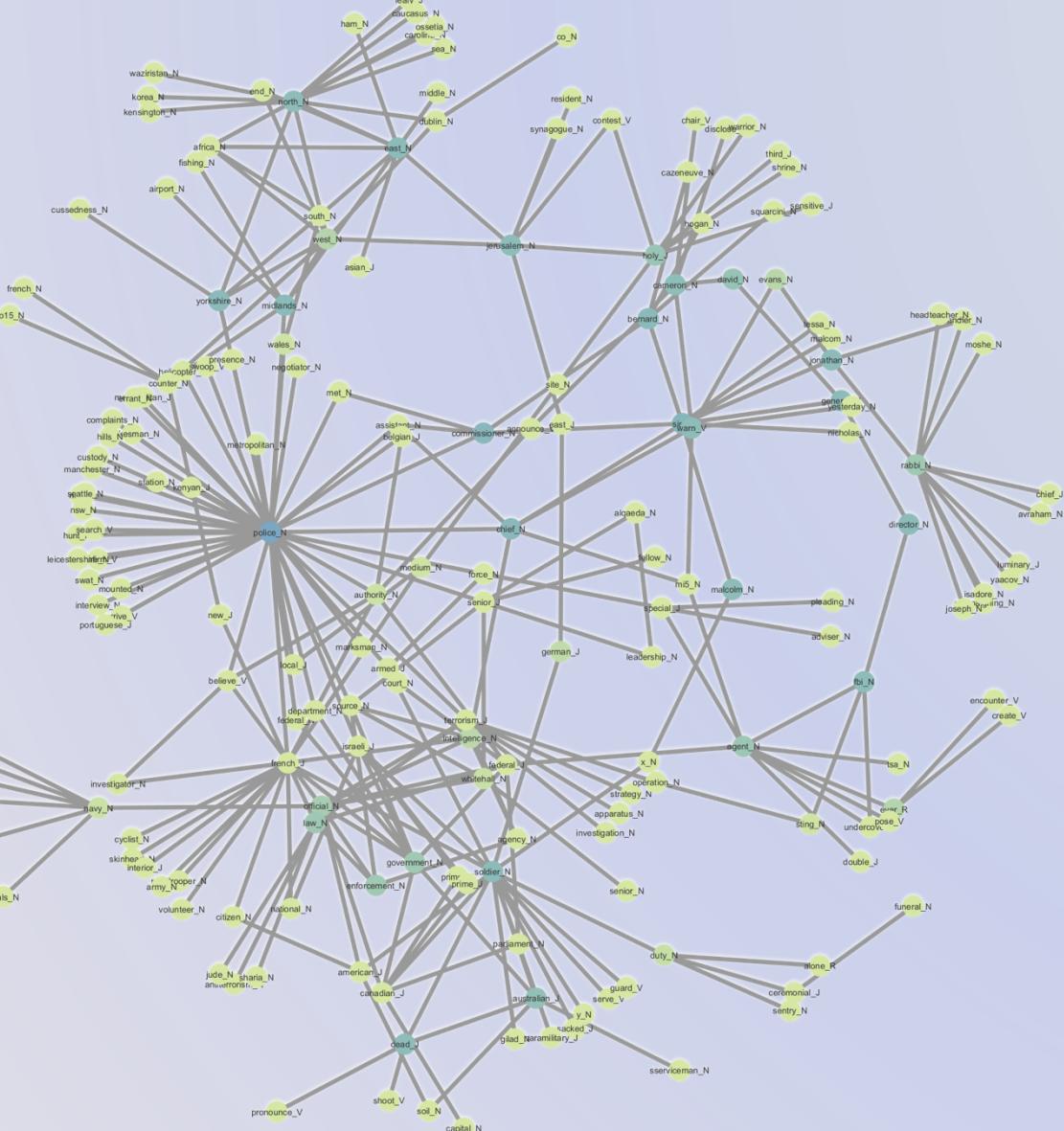


PREPEAK

POLICE & ARMY & ALQAEDA & RELIGION

HANNA SCHMÜCK - LARGE LINGUISTIC NETWORKS

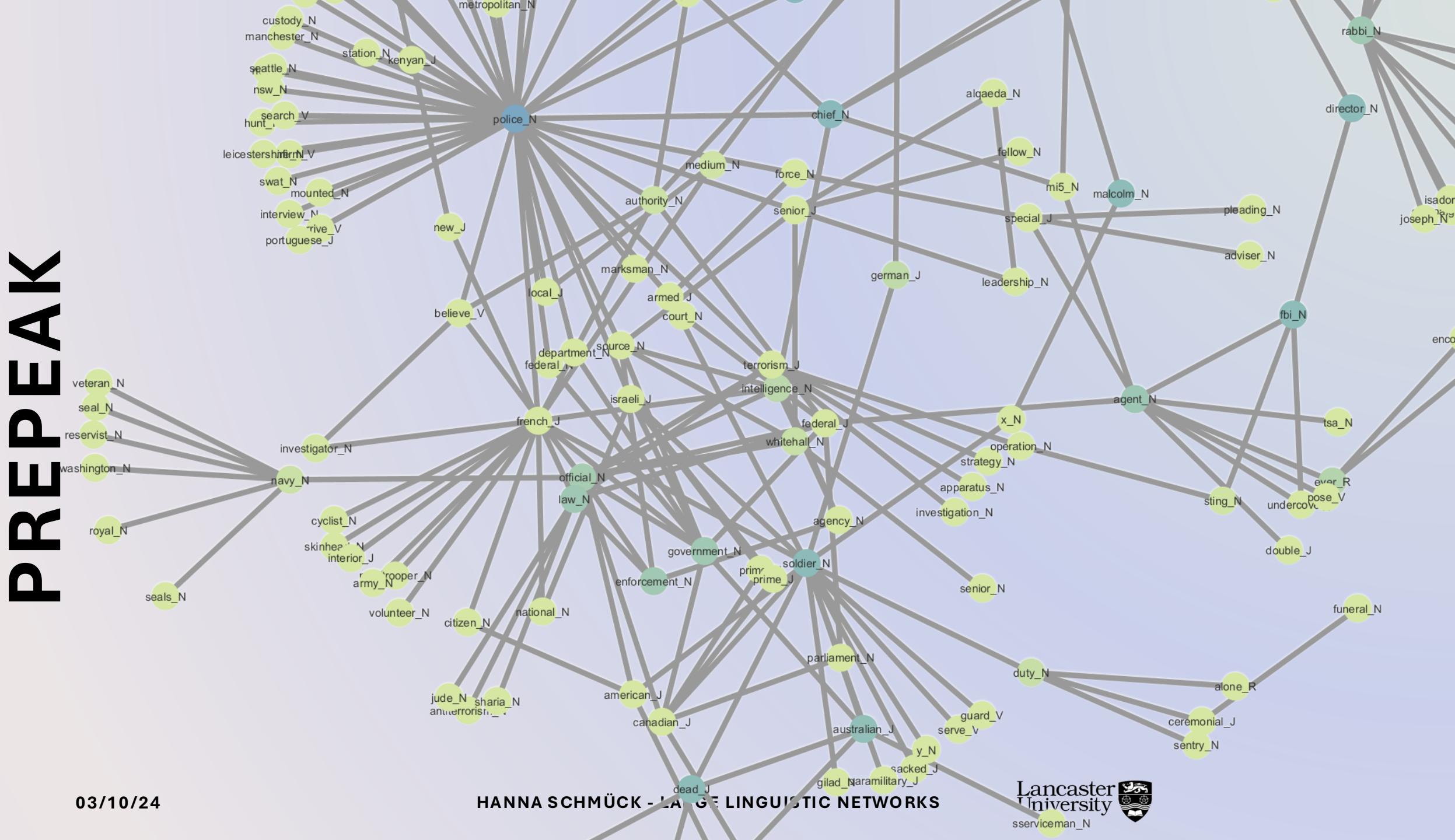
03/10/24



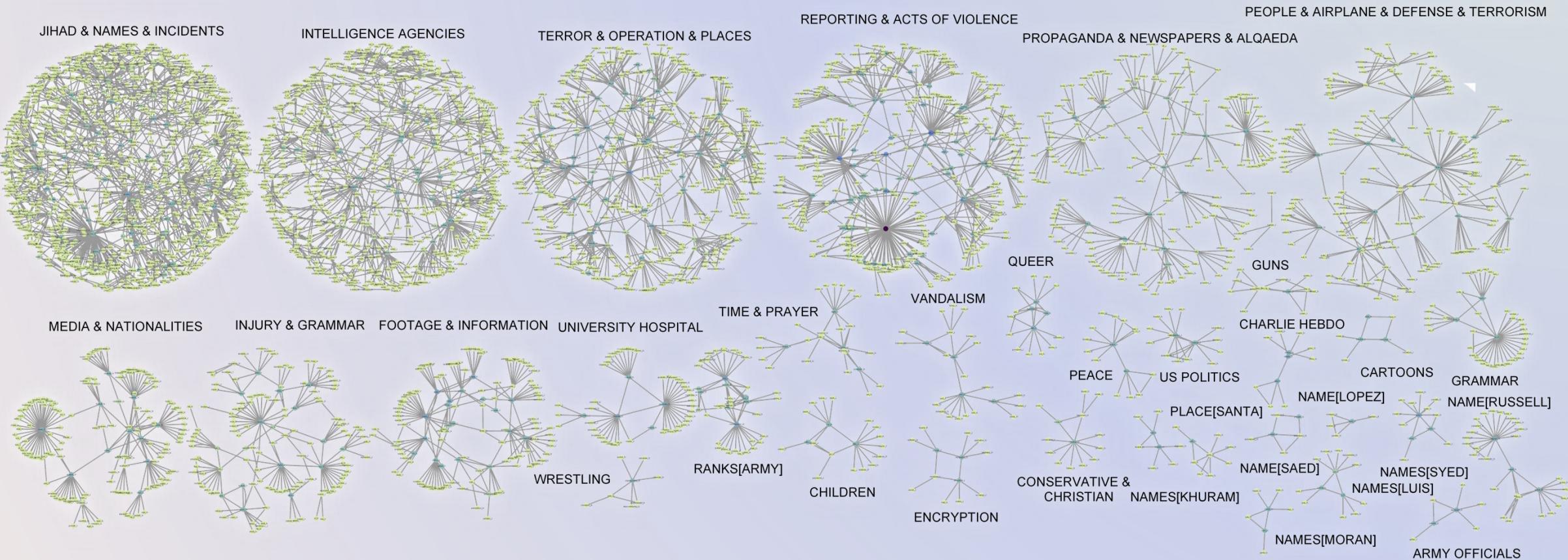
PREPEAK

03/10/24

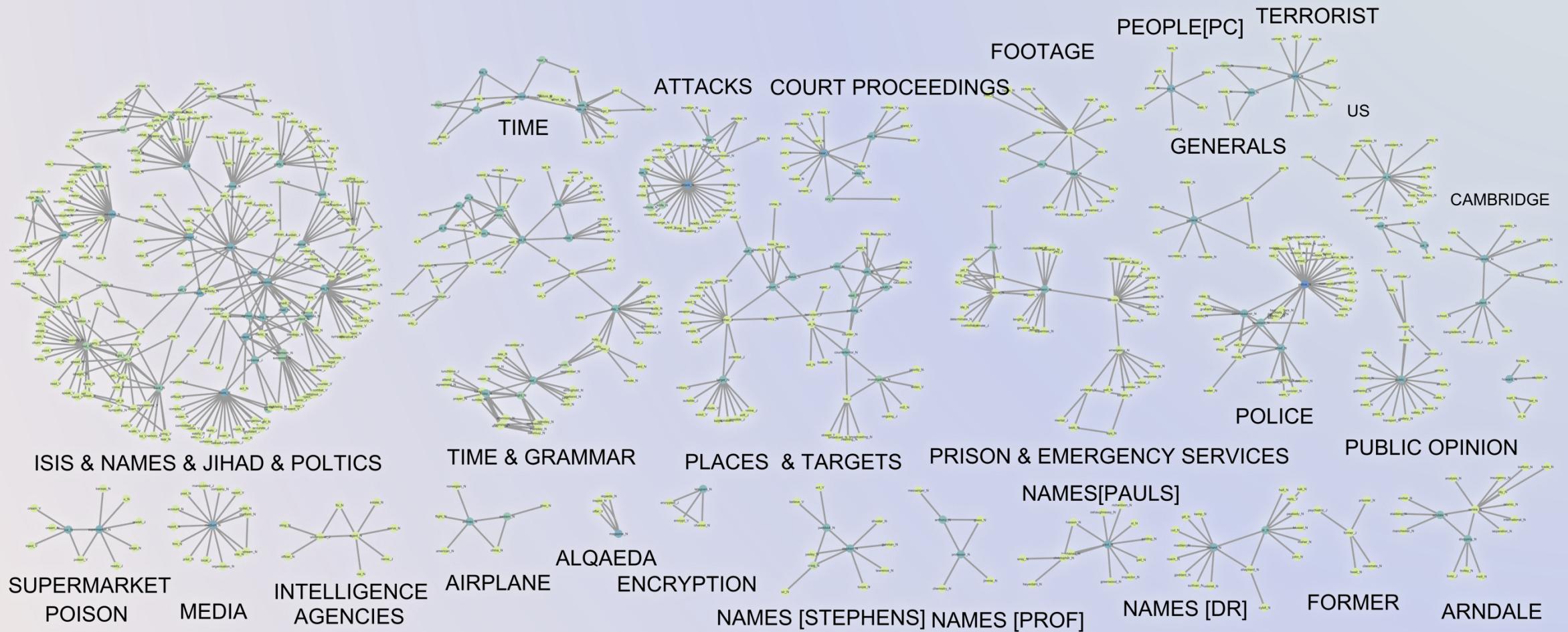
HANNA SCHMÜCK - LARGE LINGUISTIC NETWORKS



PEAK



POSTPEAK



CONCLUSION

- Reporting in the peak time is not just most verbose, but also most collocationaly dense
- Prepeak and peak are more collocationaly similar than postpeak and peak
- Overall, the collocation networks share a common core:
 - Strong focus on Islamist attacks, right wing terrorism plays a minor role
 - Prominence of organisations in the contextualisation of lone actors
- Discursive shifts:
 - Abstraction and internationalisation in the peak years
 - Loss of distinct focus on mental health (present in prepeak)

QUESTIONS?





+

THANK YOU

o

HANNA SCHMÜCK

[H.SCHMUECK@LANCASTER.AC.UK](mailto:h.schmueck@lancaster.ac.uk)

 [@HANNA_SCHMUECK](https://twitter.com/@HANNA_SCHMUECK)

 [HANNA SCHMÜCK](https://www.linkedin.com/in/hanna-schmuck/)

[HANNASCHMUECK.GITHUB.IO/](https://hannaschmueck.github.io/)

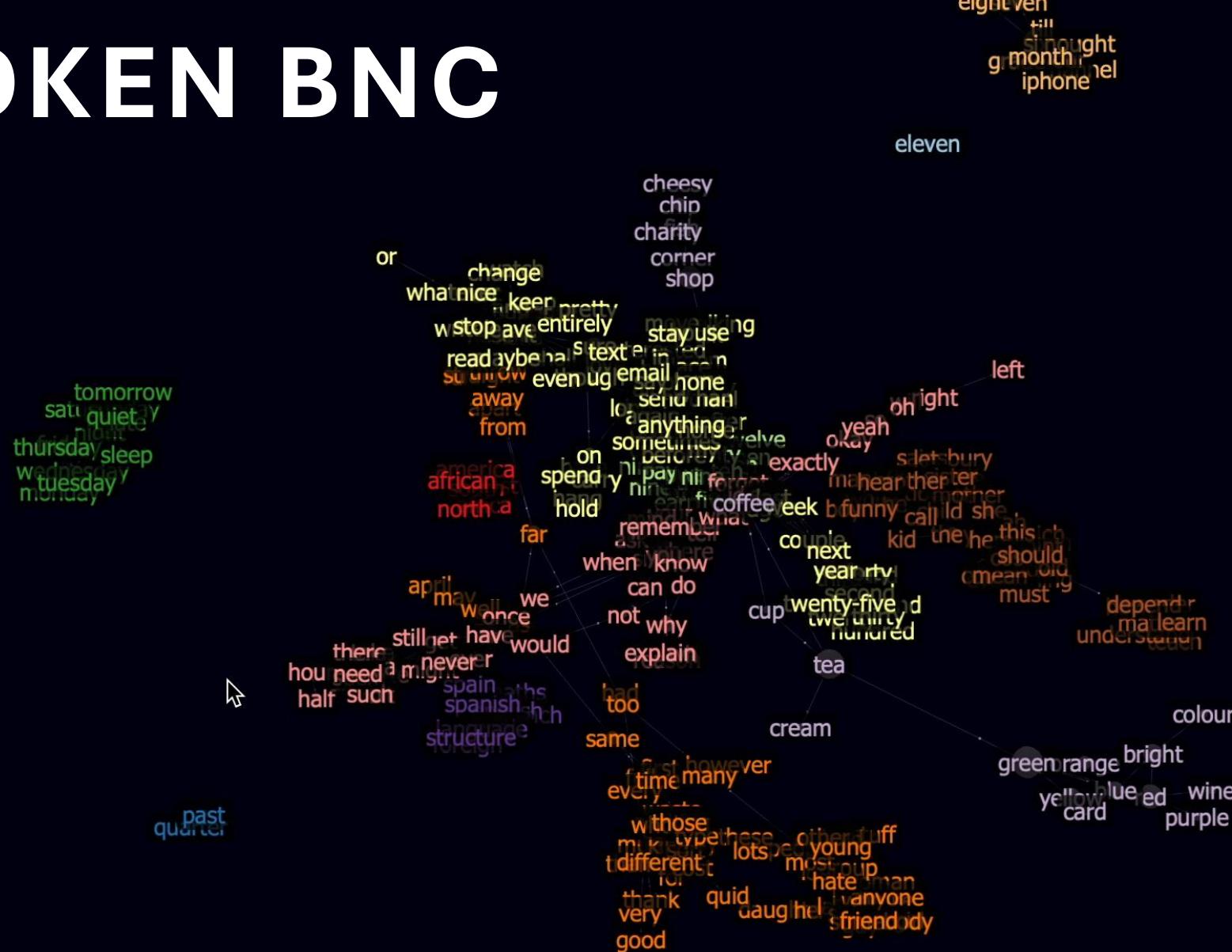
REFERENCES

- Bader, G. D., & Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(2).
<https://doi.org/10.1186/1471-2105-4-2>
- Barnbrook, G., Mason, O., & Krishnamurthy, R. (2013). The concept of collocation. In G. Barnbrook, O. Mason, & R. Krishnamurthy (Eds.), *Collocation: Applications and Implications* (1st ed., pp. 3–31). Palgrave Macmillan. https://doi.org/10.1057/9781137297242_1
- Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21(2), 139–164. <https://doi.org/10.1075/ijcl.21.2.01bak>
- Brezina, V. (2016). Collocation Networks: Exploring Associations in Discourse. In P. Baker & J. Egbert (Eds.), *Routledge Advances in Corpus Linguistics: Vol. 17. Triangulating methodological approaches in corpus-linguistic research* (pp. 90–107). Routledge.
- Brezina, V., Hawtin, A., & McEnery, T. (2021). The Written British National Corpus 2014 – design and comparability. *Text & Talk*, 41(5-6), 595–615.
<https://doi.org/10.1515/text-2020-0052>
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>
- Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox [Computer software]. <http://corpora.lancs.ac.uk/lancsbox>.
- Croft, W., & Cruse, D. A. (2004). Cognitive linguistics. Cambridge textbooks in linguistics. Cambridge University Press.
<http://www.loc.gov/catdir/description/cam032/2003053175.html>
- Gross, J. L., Yellen, J., & Anderson, M. (2019). Graph theory and its applications (Third edition). Textbooks in mathematics. CRC Press.
<https://www.taylorfrancis.com/books/mono/10.1201/9780429425134/graph-theory-applications-jonathan-gross-jay-yellen-mark-anderson>
<https://doi.org/10.1201/9780429425134>
- Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7–15. [https://doi.org/10.1016/0020-0190\(89\)90102-6](https://doi.org/10.1016/0020-0190(89)90102-6)
- Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics*, 22(3), 319–344. <https://doi.org/10.1075/ijcl.22.3.02lov>
- Malone, D. (2020). Developing a complex query to build a specialised corpus: Reducing the issue of polysemous query terms. *Corpora and Discourse International Conference 2020*. Advance online publication. <https://doi.org/10.13140/RG.2.2.31214.43846>

REFERENCES

- McEnery, T., & Brezina, V. (2019). Collocations and colligations: Visualizing lexicogrammar. In B. Busse & R. Moehlig-Falke (Eds.), Topics in English linguistics: Vol. 104. Patterns in language and linguistics: New perspectives on a ubiquitous concept (pp. 97–124). De Gruyter Mouton. <https://doi.org/10.1515/9783110596656-005>
- McEnery, T., & Hardie, A. (2011). What is corpus linguistics? *Corpus Linguistics*, 1–24. <https://doi.org/10.1017/cbo9780511981395.002>
- McGillivray, B., & Tóth, G. (2020). Collocation. In Applying Language Technology in Humanities Research. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-46493-6_4
- Noble, H., & Heale, R. (2019). Triangulation in research, with examples. *Evidence-Based Nursing*, 22(3), 67–68. <https://doi.org/10.1136/ebnurs-2019-103145>
- Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation*, 44, 137–158. <https://doi.org/10.1007/s10579-009-9101-4>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504. <https://doi.org/10.1101/gr.1239303>
- Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4), 487–512. <https://doi.org/10.1093/applin/amp058>
- Stulpinaitė, M., Horbačauskiene, J., & Kasperavičienė, R. (2016). Issues in Translation of Linguistic Collocations: Lingvistinių kolokacijų vertimo ypatumai. *Studies About Languages*(29), 31–41. <https://doi.org/10.5755/j01.sal.0.29.15056>
- Tucker, B. V., & Ernestus, M. (2016). New Questions for the Next Decade. *The Mental Lexicon*, 11(3), 375–400. <https://doi.org/10.1075/ml.11.3.03tuc>
- Wyatt, S. (2022). Interdisciplinarity: Models and Values for Digital Humanism. In H. Werthner, E. Prem, E. A. Lee, & C. Ghezzi (Eds.), Perspectives on Digital Humanism (pp. 329–334). Springer International Publishing.
- Xiao, R., & McEnery, T. (2006). Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics*, 27(1), 103–129. <https://doi.org/10.1093/applin/ami045>

SPOKEN BNC



ACADEMIC BNC



ACADEMIC BNC – SPOKEN BNC

