MAPPING MEANING: LARGE LINGUISTIC NETWORKS IN THE DIGITAL HUMANITIES LANDSCAPE

Hanna Schmück

Lancaster University

Lancaster University

CASS
Corpus Approaches to Social Science

TOPICS

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…

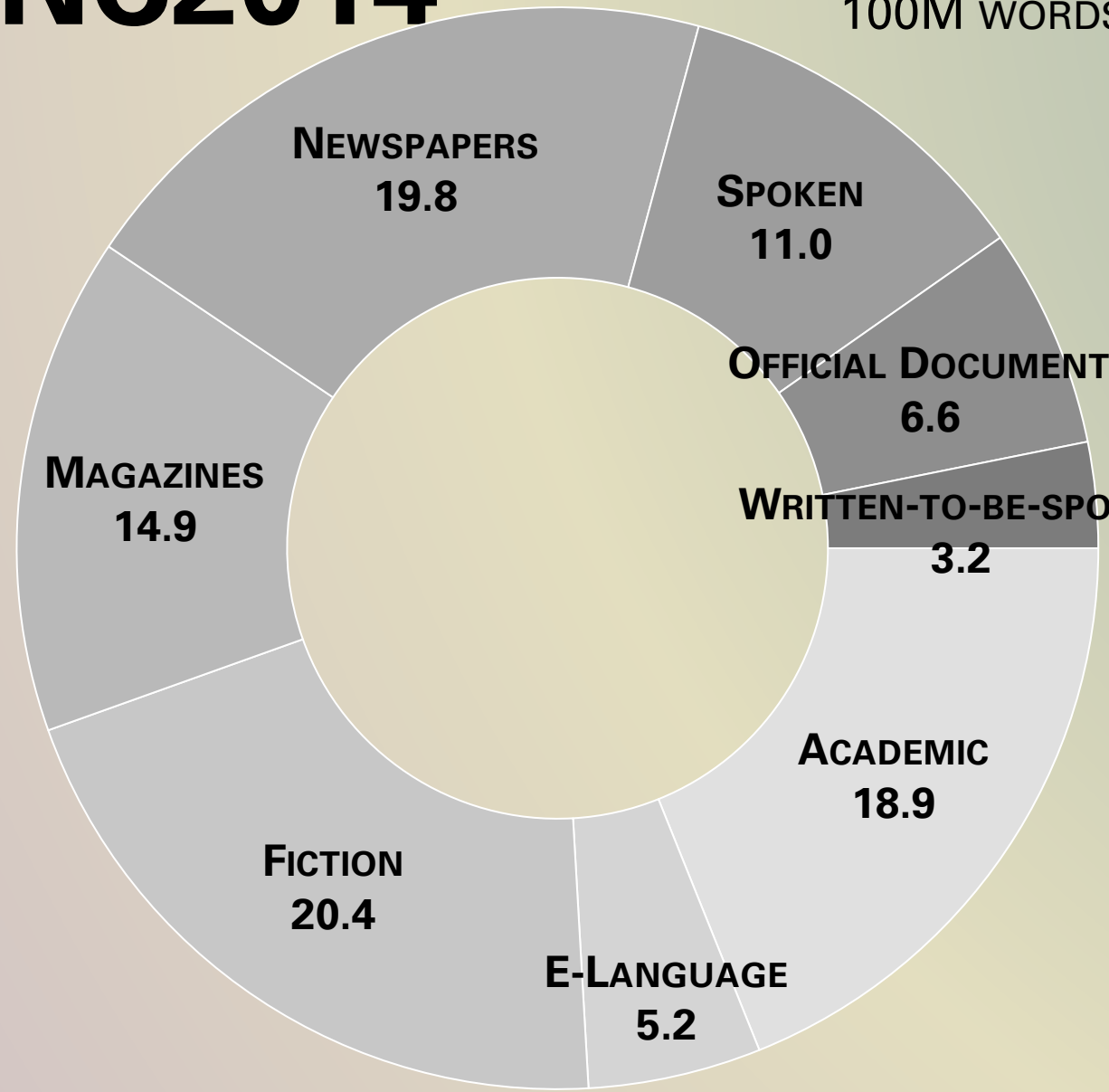
- A large body of text
- Aims to be representative of language (or a genre of language)
- Machine readable (McEnery & Hardie, 2011)

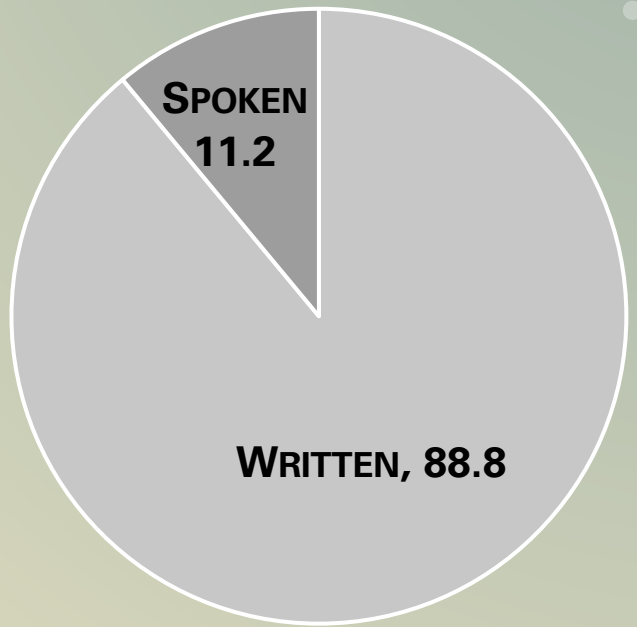
Often annotated with additional linguistic information – e.g. grammatical codes, learner errors etc.

THE BNC2014

100M WORDS OF CONTEMPORARY BRITISH ENGLISH

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





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.

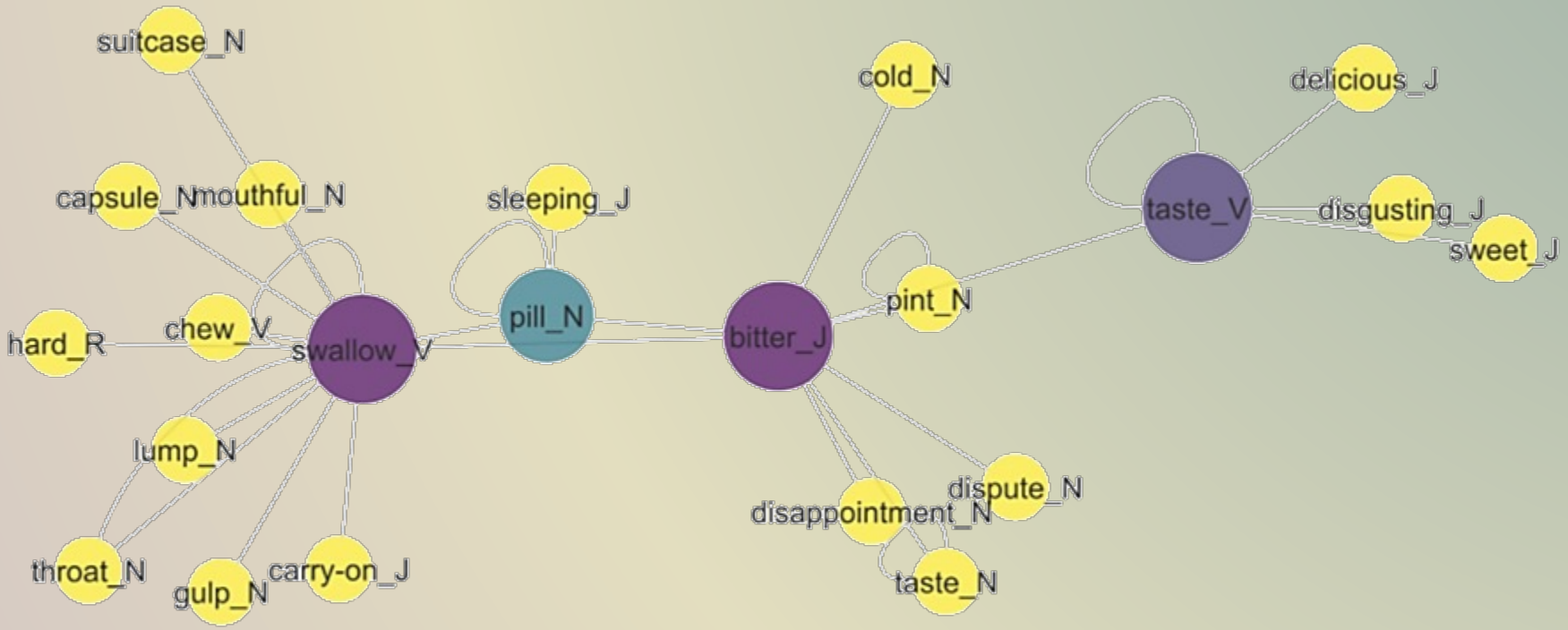
WHAT IS A COLLOCATION?

- 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)

EXAMPLE



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

WHY USE COLLOCATION NETWORKS?

VISUALISATION OF

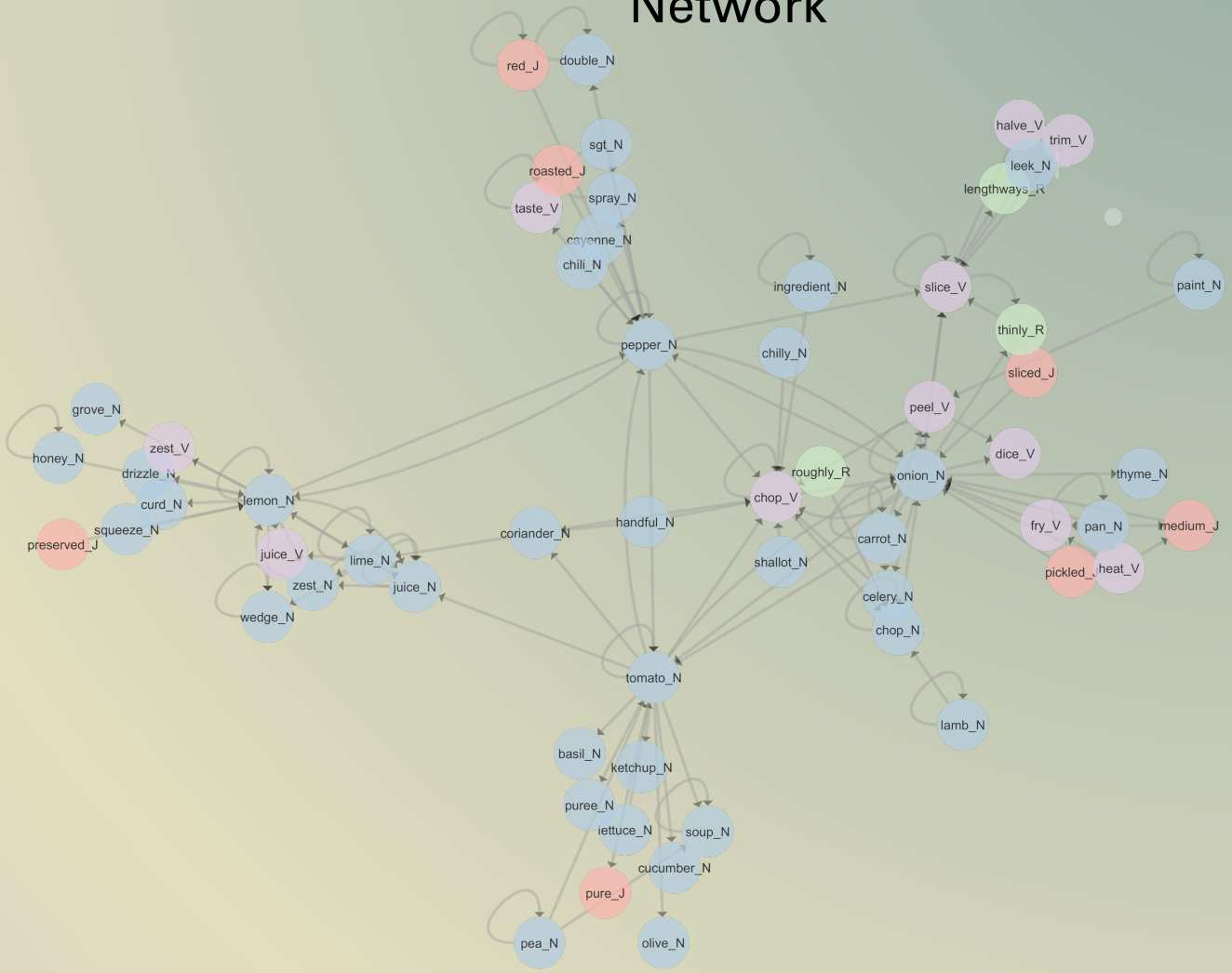
- **DISCOURSE RELATIONSHIPS, ABOUTNESS** OF A TEXT OR DISCOURSE
- **SEMANTIC RELATIONS** (PECINA, 2010; XIAO & MCENERY, 2006; BREZINA, 2016; BAKER, 2016; BREZINA ET AL., 2015)
- **LEXICOGRAMMATICAL FEATURES** (MCENERY & BREZINA, 2019)

NETWORKS VS. TABLES

Table

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			

Network



(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

Adaptability of tables

- Sorting (highest/lowest AM)
- Inclusion of more metadata/annotation

Adaptability of networks

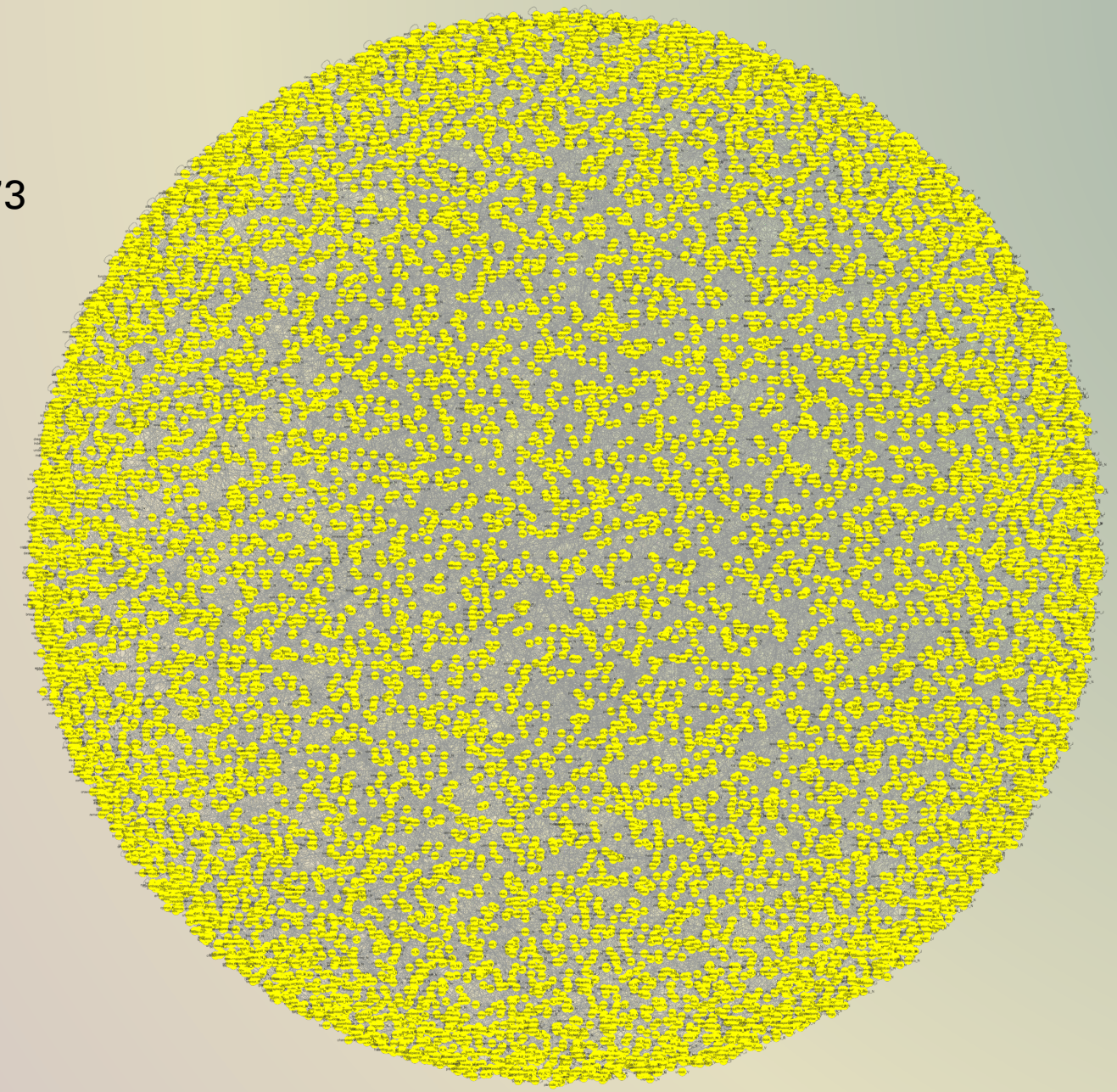
- Size
- Shape
- Colour
- Labels

LLN

BNC-2014 – logDice > 6.73

LARGEST CONNECTED COMPONENT

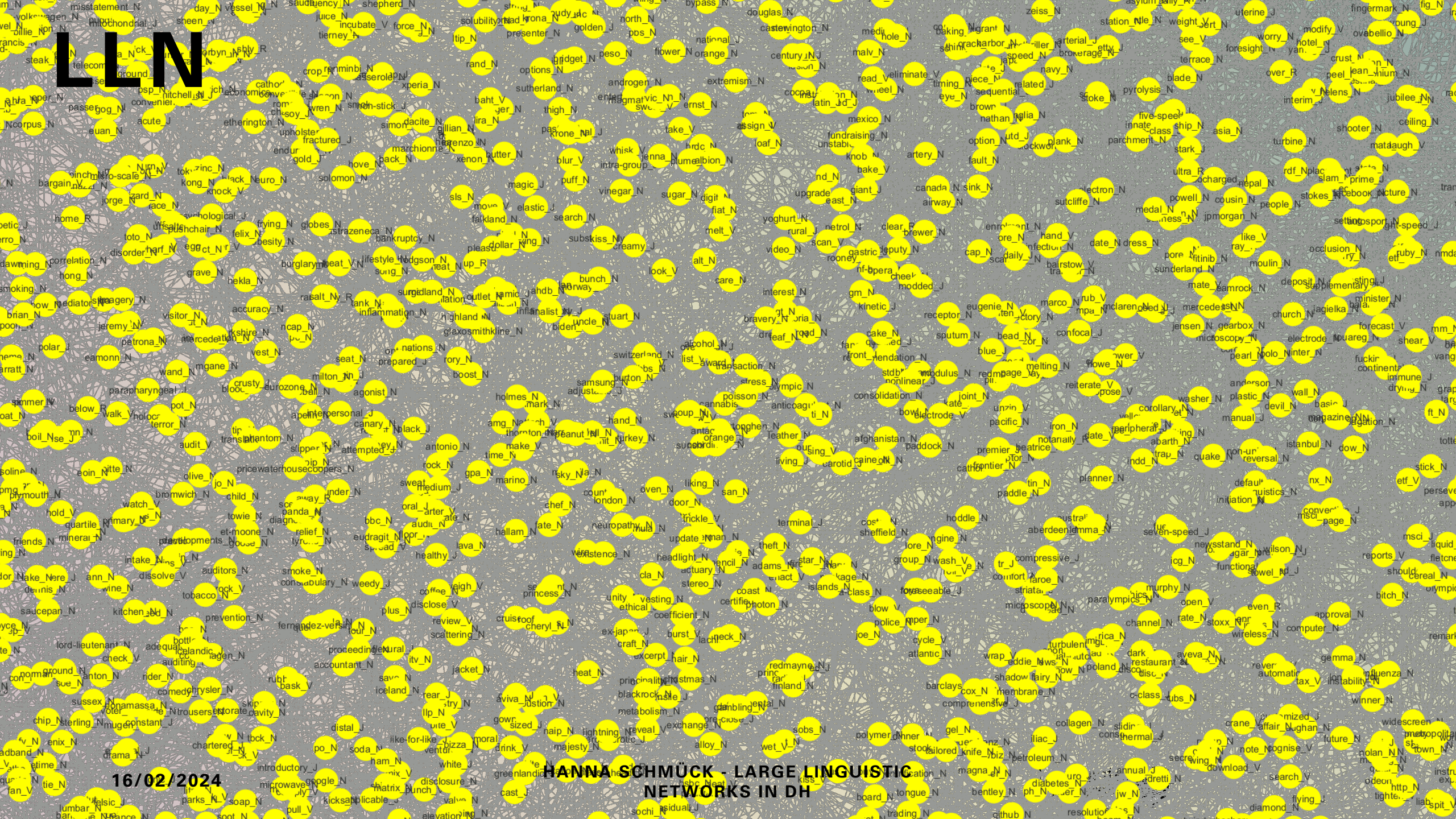
NODE COLOUR = POS



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

-

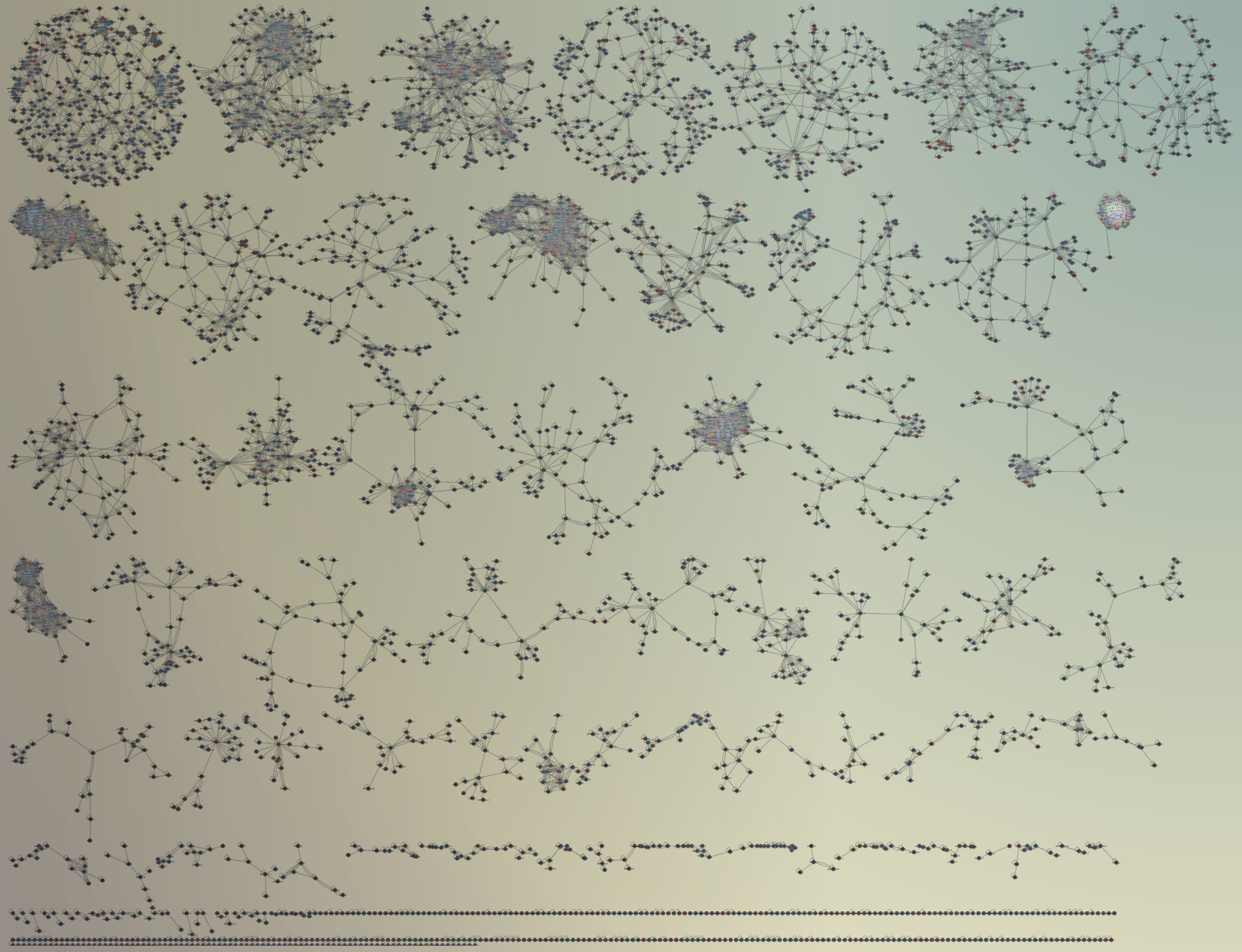
2-dimensional, Edge length: AM, Colour-coding: POS, Layout type: Edge-weighted spring embedded


LLN
16/02/2024
HANNA SCHMÜCK - LARGE LINGUISTIC NETWORKS IN DH

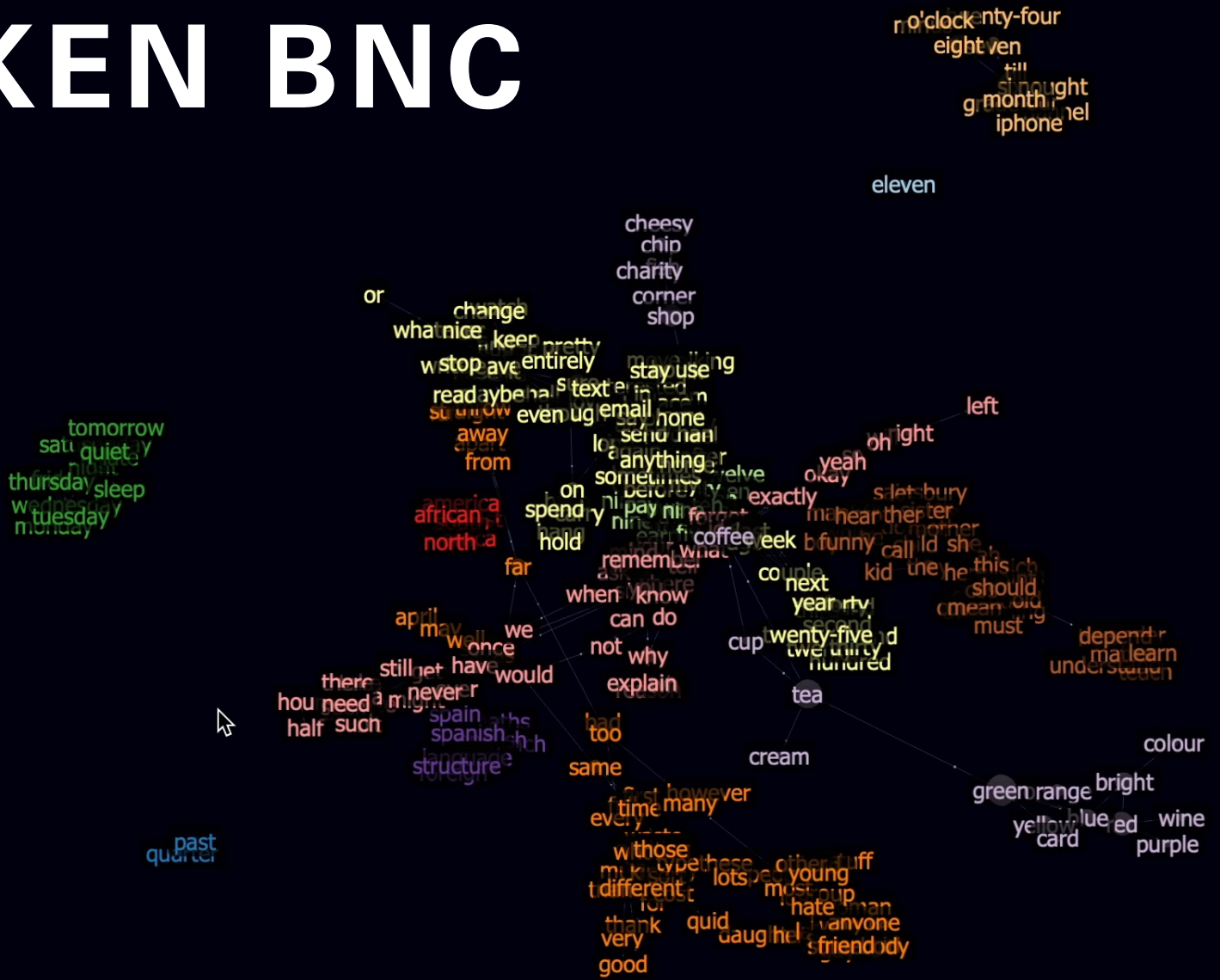
LLN

Clusters emerging from collocates of drug*|drink*|alcohol*|abuse* in the BNC2014

SAME PARAMETERS



SPOKEN BNC

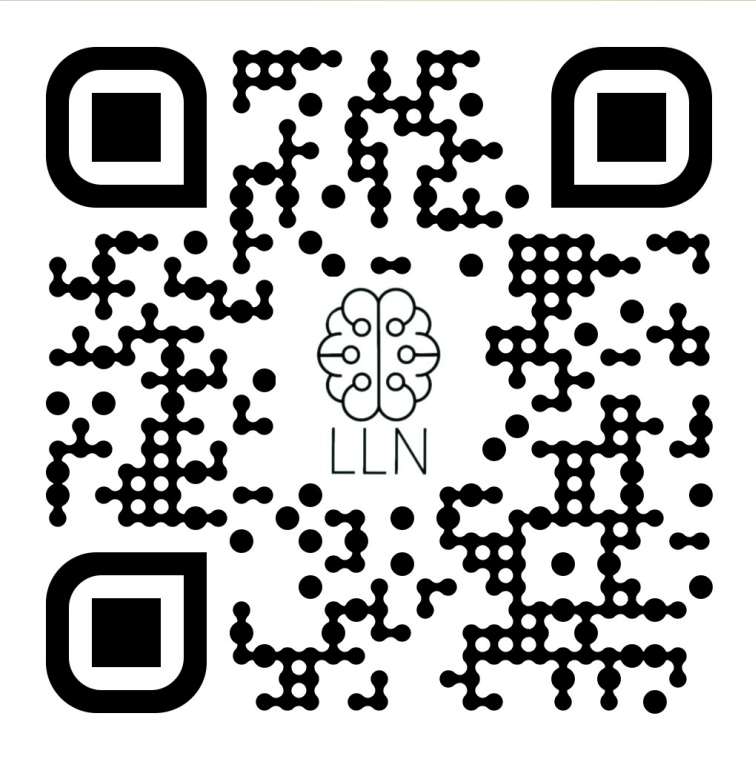


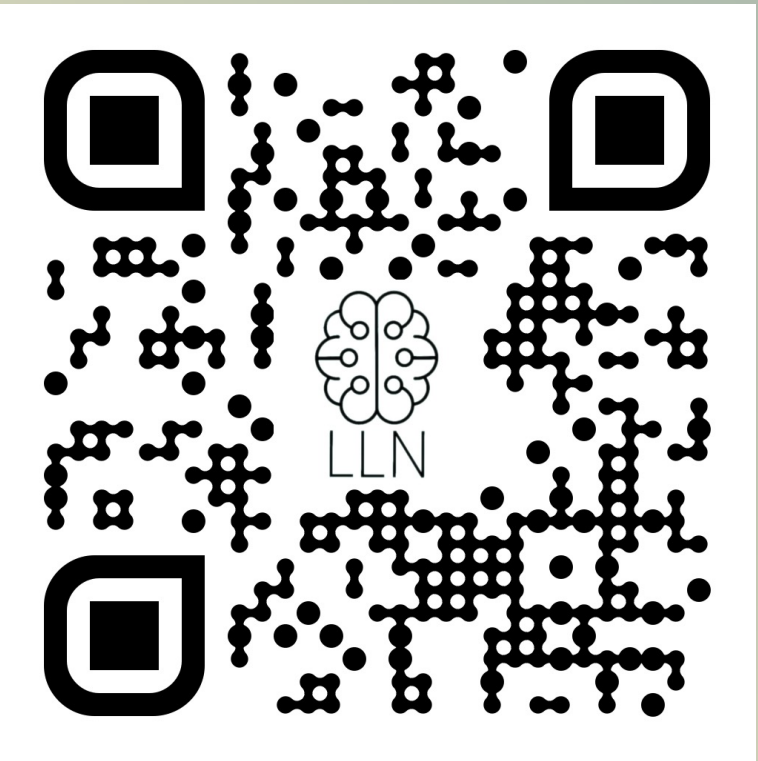
Left-click: rotate, Mouse-wheel/middle-click: zoom, Right-click: pan

ACADEMIC BNC

Left-click: rotate, Mouse-wheel/middle-click: zoom, Right-click: pan

ACADEMIC BNC – SPOKEN BNC





RELATED WORK

- Included in the PhD:
  - Systematic network comparisons – e.g. of word association networks and collocation networks
  - Integration of other datasets
- Exploring changing networks, e.g. displaying results from diachronic studies
- Including more annotation in visualisations – examples from corpora, metadata
- Allowing researchers to dynamically change existing visualisations

QUESTIONS?

THANK YOU

Hanna Schmück

h.schmueck@lancaster.ac.uk

@hanna_schmueck

Hanna Schmück

hannaschmueck.github.io/

REFERENCES

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

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

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

REFERENCES

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