Scatterchron:

Visualizing Language Change in Diachronic Corpora

Code for the interactive visualizations:
https://github.com/JasonKessler/KeynessToolsTalk/blob/main/BBC%20News%20One%20Year.ipynb

Repo:
https://github.com/JasonKessler/KeynessToolsTalk

Jason S. Kessler || Using and Developing Software for Keyness Analysis || February 27, 2023

Introduction

- Why would we like to visualize how language changes over time?
- **Temporal.** News alerts are ordered by date.
  - Create a timeline of events over a year.
- **Sequential.** Chapters or pages in a novel.
  - Visually represent the plot and presence of characters.
- **Product review data.** Reviews are ordered by star rating.
  - See what features or issues are in different rating regions.

Outline

- Related work in visualizing language change over time
- Interactive text visualizations with Scattertext
- Binary difference: term frequencies before and after time-change
- Most associated terms in each time step.
  - Clustering time steps to make them more manageable
- Detecting terms that clump together at different time steps or are more dispersed
- Average time position
- Putting it all together: **Scatterchron**

Related Work

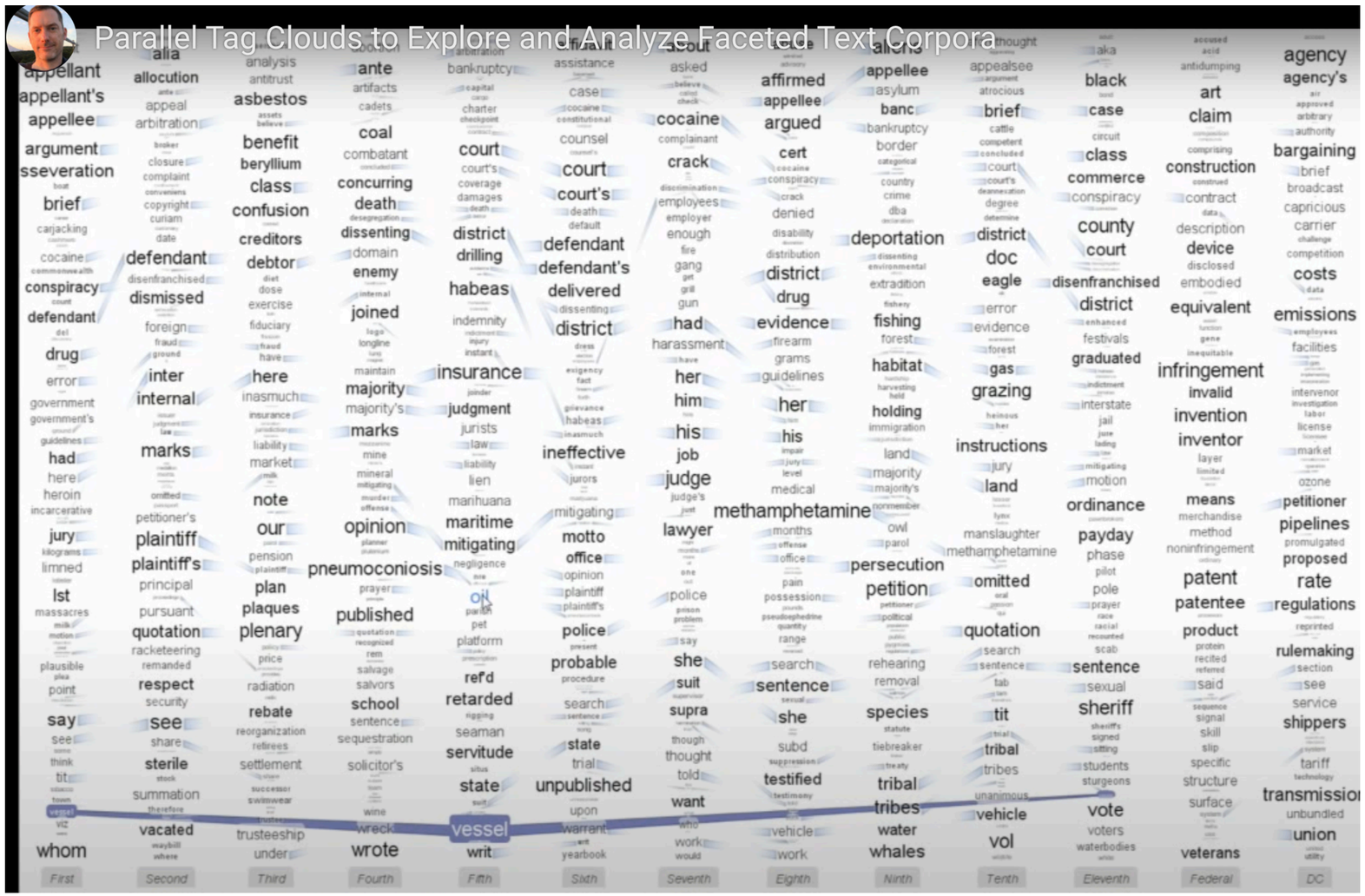
Parallel Tag Clouds

Key terms are vertically arranged at each time step

Size indicates time-step association (G^2)

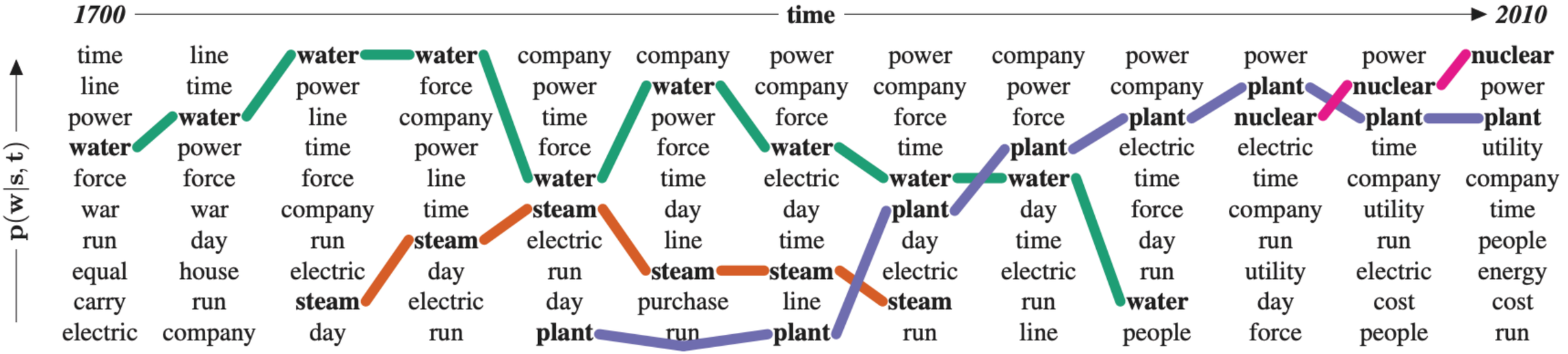
Order is alphabetical

Connectors to the same words appearing in different time steps



C. Collins, F. B. Viégas, and M. Wattenberg. 2009. Parallel Tag Clouds to Explore and Analyze Facted Text Corpora. VAST.

Related Work



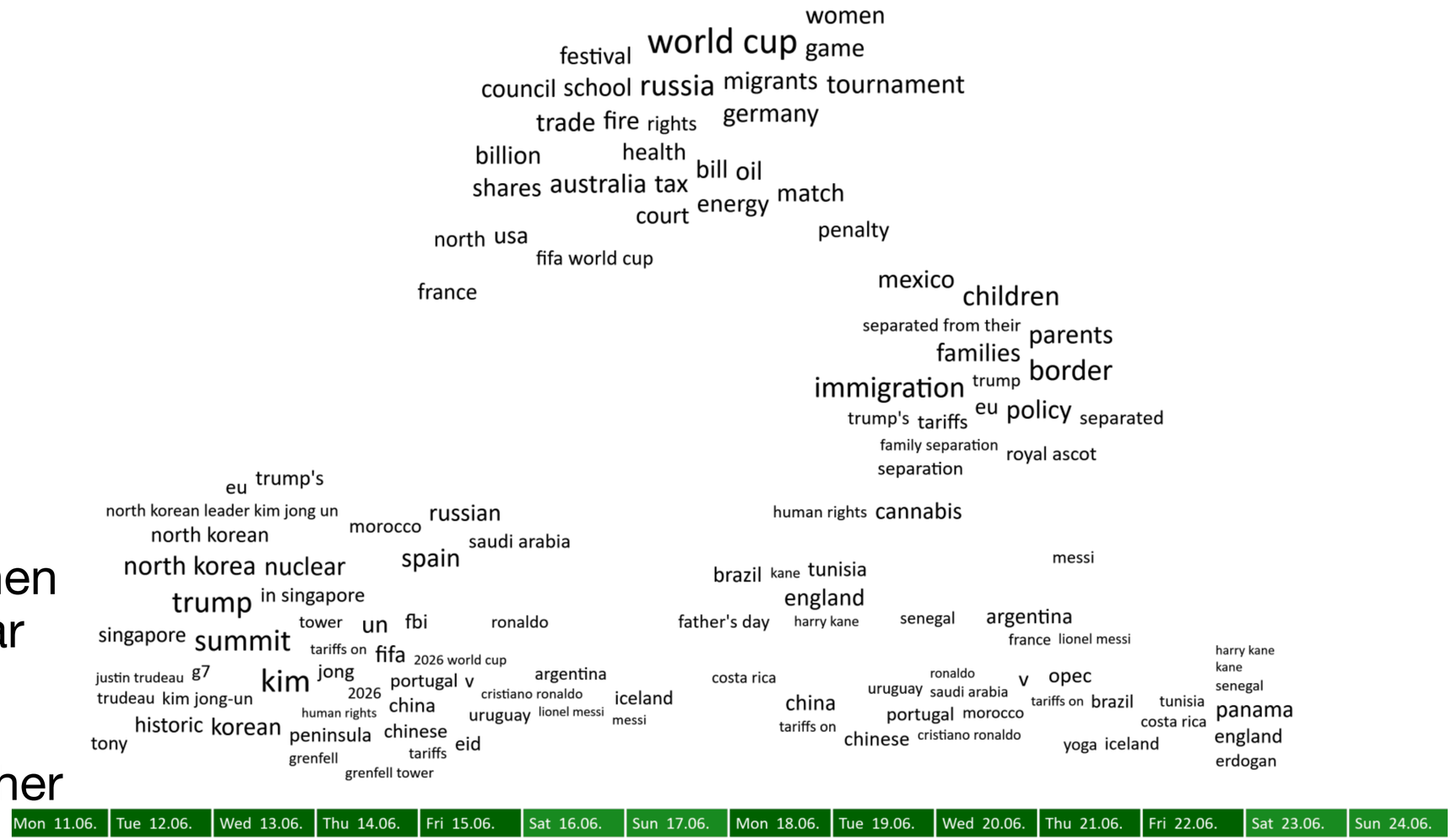
- “Words most associated with the energy-sense of the word ‘power’”
- Influenced by Parallel Tag Clouds
- More readable layout
- Ordered by association, not alphabetically

Lea Frermann and Mirella Lapata. A Bayesian Model of Diachronic Meaning Change. TACL 2016.

Related Work

PyramidTags

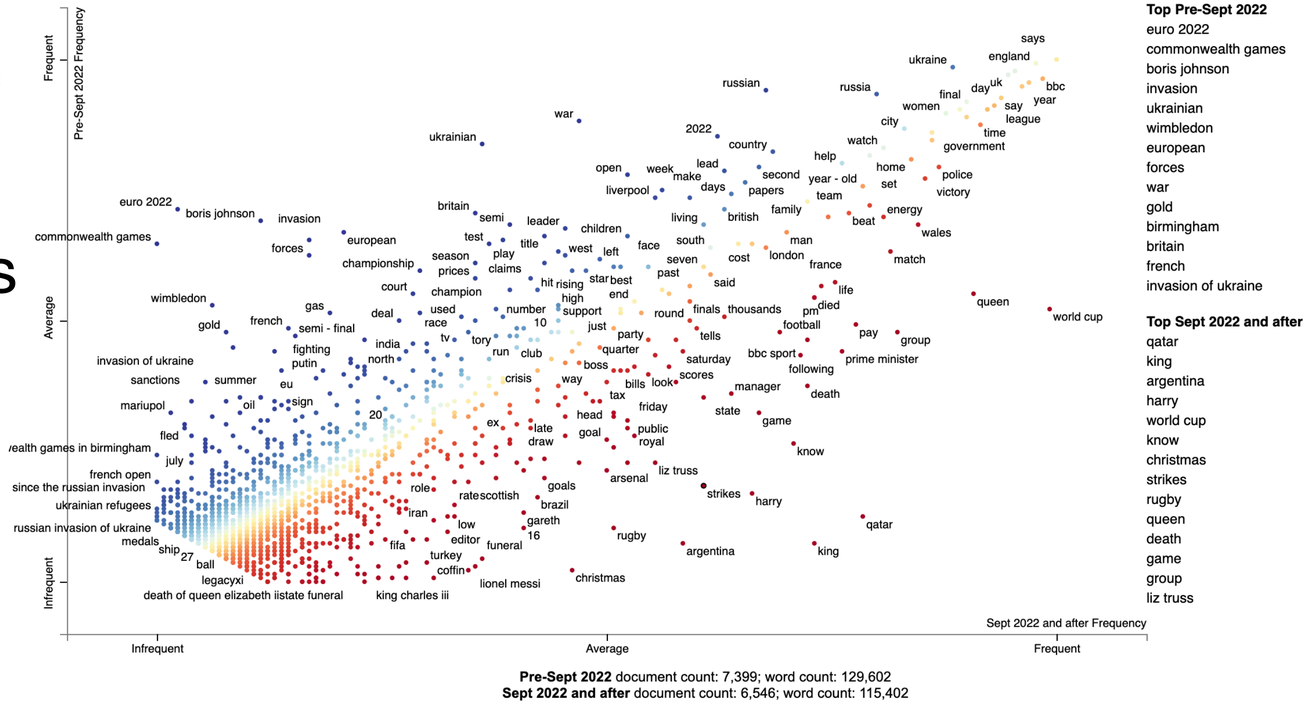
- N-grams are displayed in a “cloud” above a timeline.
- Horizontal position: roughly mean point in the timeline
- Vertical: higher up indicates longer time-span
- Arranged s.t. words, even when not in phrases, tend to appear in sequence.
  - Similar terms cluster together



Knittel, J., Koch, S., & Ertl, T. 2021. PyramidTags: Context-, Time- and Word Order-Aware Tag Maps to Explore Large Document Collections. IEEE Transactions on Visualization and Computer Graphics.

Time-step divided plot

- BBC RSS feed corpus (Feb 2022 - Feb 2023)[1]
- 13,945 alerts; 361 days; 281,589 words
- Label each alert by before Sept 2022, Sept 2022 and after
- Plot rank frequency of features on x-axis in <9/22 and of >9/22 on y-axis
- Uses Scattertext (Kessler 2017)
- Color by difference in ranked frequency



Pre-Sept 2022 document count: 7,399; word count: 129,602
Sept 2022 and after document count: 6,546; word count: 115,402

https://jasonkessler.github.io/keynesstalk/bbc_divided.html

Jason S. Kessler. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. ACL System Demonstrations. Vancouver, BC. 2017

1. kaggle.com/code/gpreda/bbc-news-rss-feeds

Parallel tag clouds

Displaying time-step information

2022-02-22	2022-02-24	2022-02-26	2022-02-28	2022-03-01	2022-03-02	2022-03-03	2022-03-04	2022-03-05	2022-03-06	2022-03-07
restrictions	covid	watched	spread	rules	care	crime	russia	russia	war	ukraine
announced	people	virus	virus	face	fourth	really	ukraine	resistance	ukrainian	russian
covid	england	needs	key	uk	covid	tackle	experts	happen	manchester	woman
end	self	covid	sector	pitch	workers	little	reasons	warned	investigated	crew
government	tests	living	unlikely	carrying	people	economic	coronavirus in the uk	president putin	rival	russia
england	longer	uk	long - term	alice	england	putin	explore the data on coronavirus	answers	city	irish
striking images from our readers	free	canadian	beach	peace	thrown	late	relatives	kevin	parents	19
ms	government	offers	mercedes	seal	broadcaster	money	offensive	missile	russian	shock
unusual	million people have fled	holding	threatening	easy	waste	ukraine says	items	jeremy	national	leave
readers around the world	followed	people have fled their homes	racing	golden	alice	trains	feel	really	qatar	year - old

- “covid” is moused-over
- Jenson Shannon Divergence is used as time-step association metric
- Symmetric version of KL divergence
- This approximates Cohen’s d well and is faster to calculate

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. PloS one, 10(10):e0137041.

Parallel tag clouds

Displaying time-step information

2022-02-22	2022-02-24	2022-02-26	2022-02-28	2022-03-01	2022-03-02	2022-03-03	2022-03-04	2022-03-05	2022-03-06	2022-03-07
restrictions	covid	watched	spread	rules	care	crime	russia	russia	war	ukraine
announced	people	virus	virus	face	fourth	really	ukraine	resistance	ukrainian	russian
covid	england	needs	key	uk	covid	tackle	experts	happen	manchester	woman
end	self	covid	sector	pitch	workers	little	reasons	warned	investigated	crew
government	tests	living	unlikely	carrying	people	economic	coronavirus in the uk	president putin	rival	russia
england	longer	uk	long - term	alice	england	putin	explore the data on coronavirus	answers	city	irish
striking images from our readers	free	canadian	beach	peace	thrown	late	relatives	kevin	parents	19
ms	government	offers	mercedes	seal	broadcaster	money	offensive	missile	russian	shock
unusual	million people have fled	holding	threatening	easy	waste	ukraine says	items	jeremy	national	leave
readers around the world	followed	people have fled their homes	racing	golden	alice	trains	feel	really	qatar	year - old

- Size: frequency in class
- Ordered by delta JSD; tends to choose frequently but associated terms
- Problem: there are 361 time steps; we can’t show them all without a lot of horizontal scrolling

Interactive visualization: https://jasonkessler.github.io/keynesstalk/bbc_dispersion_table.html

Clustered parallel tag clouds

Displaying time-step information compactly

2022-02-22 to 2022-07-30	2022-07-31 to 2022-08-16	2022-08-17 to 2022-08-25	2022-08-26 to 2022-08-27	2022-08-28 to 2022-09-15	2022-09-16 to 2023-02-20
ukraine	commonwealth games	european	energy	queen	world cup
russian	gold	championships	ben	king	bbc
russia	birmingham	students	bills	coffin	qatar
war	england	britain	ben stokes	charles	argentina
ukrainian	commonwealth games in birmingham	european championships	days	queen elizabeth ii	christmas
boris johnson	wins	results	way	liz truss	group
euro 2022	2022	south	title	death	strikes
invasion	women	manchester	people	king charles iii	harry
city	win	great britain	rivals	state	rugby
wimbledon	watch	test	man	queen 's coffin	game

- Solution: create clusters of time steps
- Among terms displayed, compute Spearman’s correlation coefficient on score ranks for adjacent time steps
- Select the K (above K=5) lowest correlations for breakpoints

Plotting change over time: dispersion

- Dispersion: metric which indicates if terms appear at the same frequency distributed across corpus parts (time steps) or if they occur with disproportionate frequency in some rather than others
- Dispersion metric: DA (Burch et al., 2017)
- Mean absolute pairwise difference between part frequencies
- Ranges from 0 (a term only appears in one part) to 1 (same frequency in each part)
  - 0.5: the average pairwise difference is the mean frequency
- It takes O(time-steps^2) to run. Gries' DP, etc. are faster but lower quality.

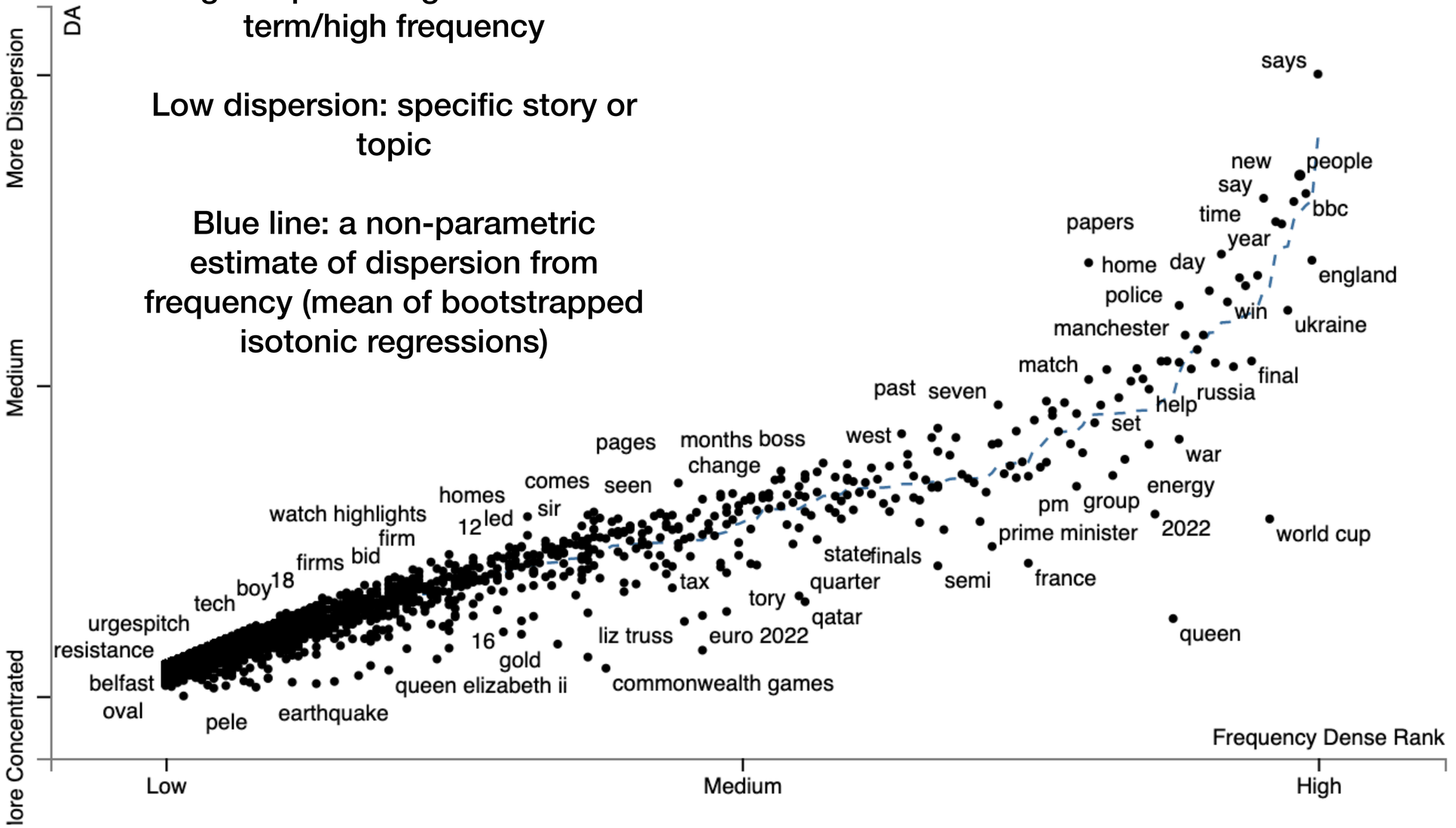
Burch, B., Egbert, J., & Biber, D. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*

Plotting change over time: dispersion

High dispersion: general news term/high frequency

Low dispersion: specific story or topic

Blue line: a non-parametric estimate of dispersion from frequency (mean of bootstrapped isotonic regressions)



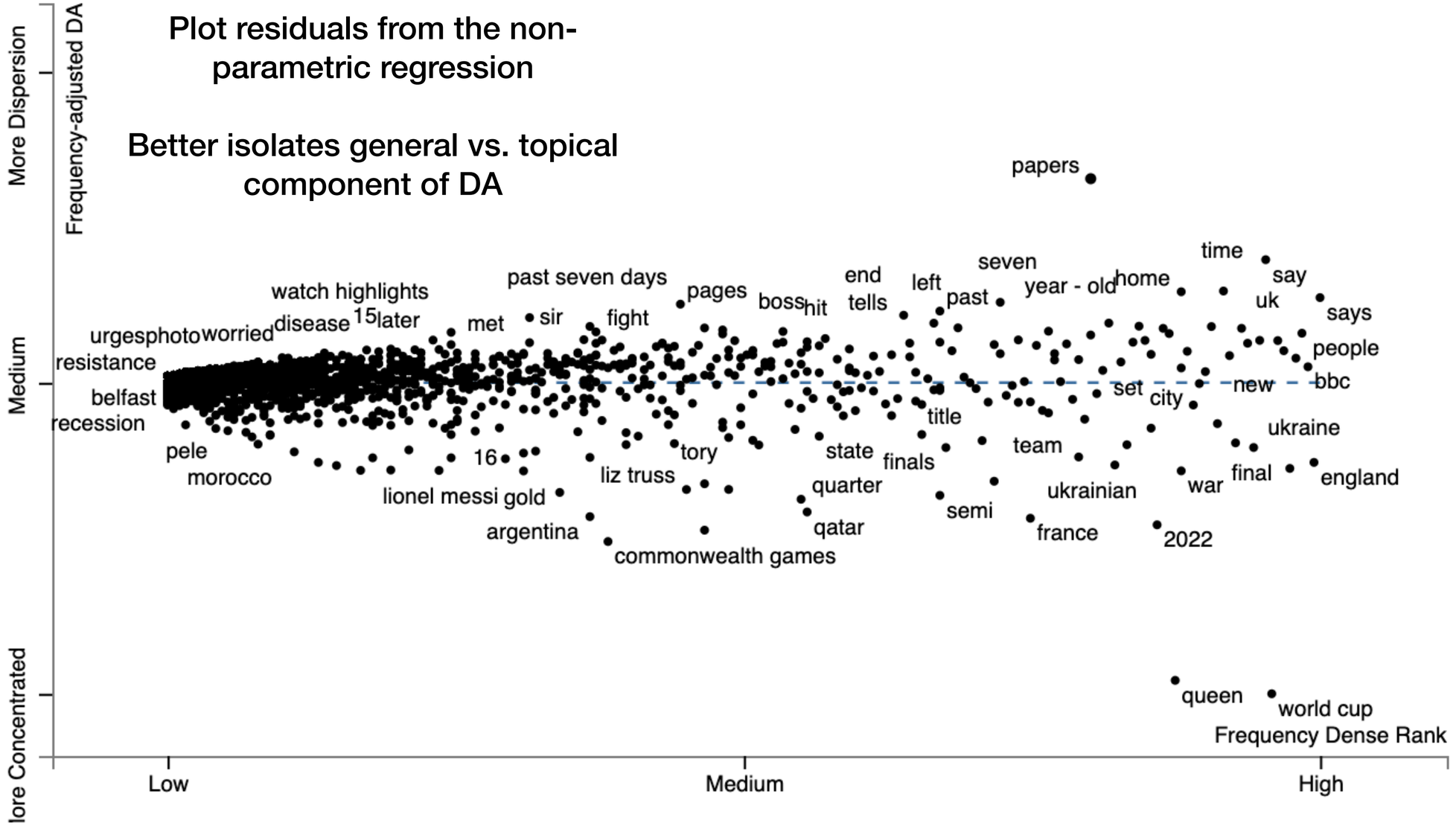
Dispersed

says
people
bbc
say
new
uk
year
time

Concentrated

pele
world cup quarter - final
earthquake
european championship
senegal
2022 commonwealth ga
queen 's coffin
oval

Residual DA: frequency normalization



Dispersed

papers
say
time
home
says
seven
pages
past

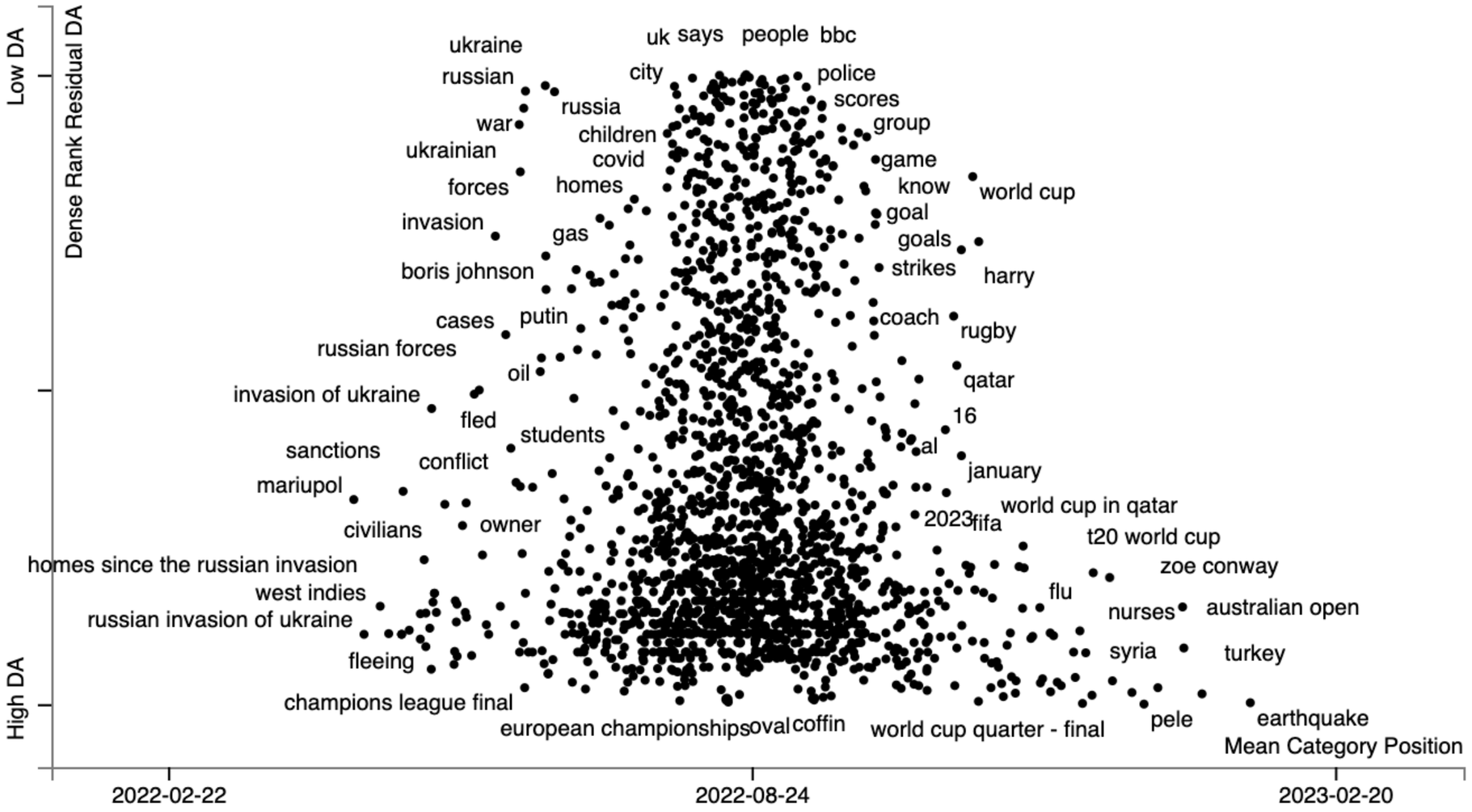
Concentrated

world cup
queen
commonwealth games
euro 2022
2022
france
argentina
qatar

Incorporating Average Position

- Residual dispersion: what’s event-related or bursty vs. what’s more general
- Missing: where on the timeline did the event occur
- Solution: center terms on their average time steps, similar to PyramidTags

Incorporating Average Position



Most Dispersed

says
people
bbc
say
new
uk
year
time

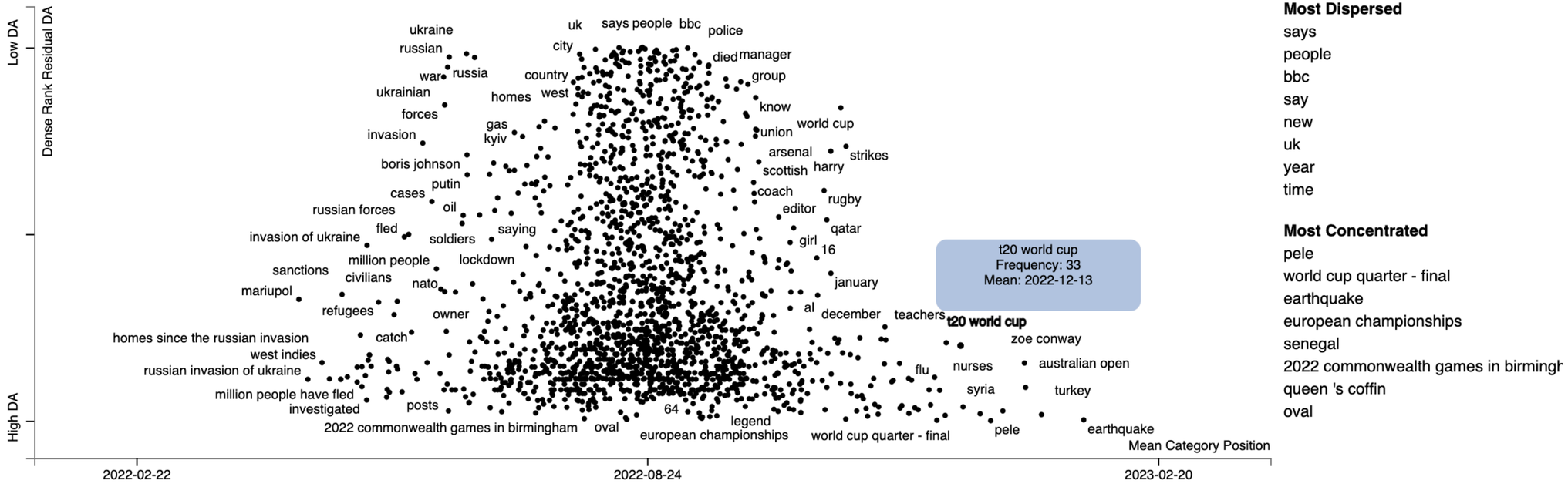
Most Concentrated

pele
world cup quarter - final
earthquake
european championships
senegal
2022 commonwealth games in birmingh
queen 's coffin
oval

Scatterchron: stacking the visualizations

2022-02-22 to 2022-07-30	2022-07-31 to 2022-08-16	2022-08-17 to 2022-08-25	2022-08-26 to 2022-08-27	2022-08-28 to 2022-09-15	2022-09-16 to 2023-02-20
ukraine	commonwealth games	european	energy	queen	world cup
russian	gold	championships	ben	king	bbc
russia	birmingham	students	bills	coffin	qatar
war	england	britain	ben stokes	charles	argentina
ukrainian	commonwealth games in birmingham	european championships	days	queen elizabeth ii	christmas
boris johnson	wins	results	way	liz truss	group
euro 2022	2022	south	title	death	strikes
invasion	women	manchester	people	king charles iii	harry
city	win	great britain	rivals	state	rugby
wimbledon	watch	test	man	queen 's coffin	game

Interactive plot: https://jasonkessler.github.io/keynesstalk/bbc_grouped_time_table.html



Most Dispersed

says
people
bbc
say
new
uk
year
time

Most Concentrated

pele
world cup quarter - final
earthquake
european championships
senegal
2022 commonwealth games in birmingh
queen 's coffin
oval

Future work

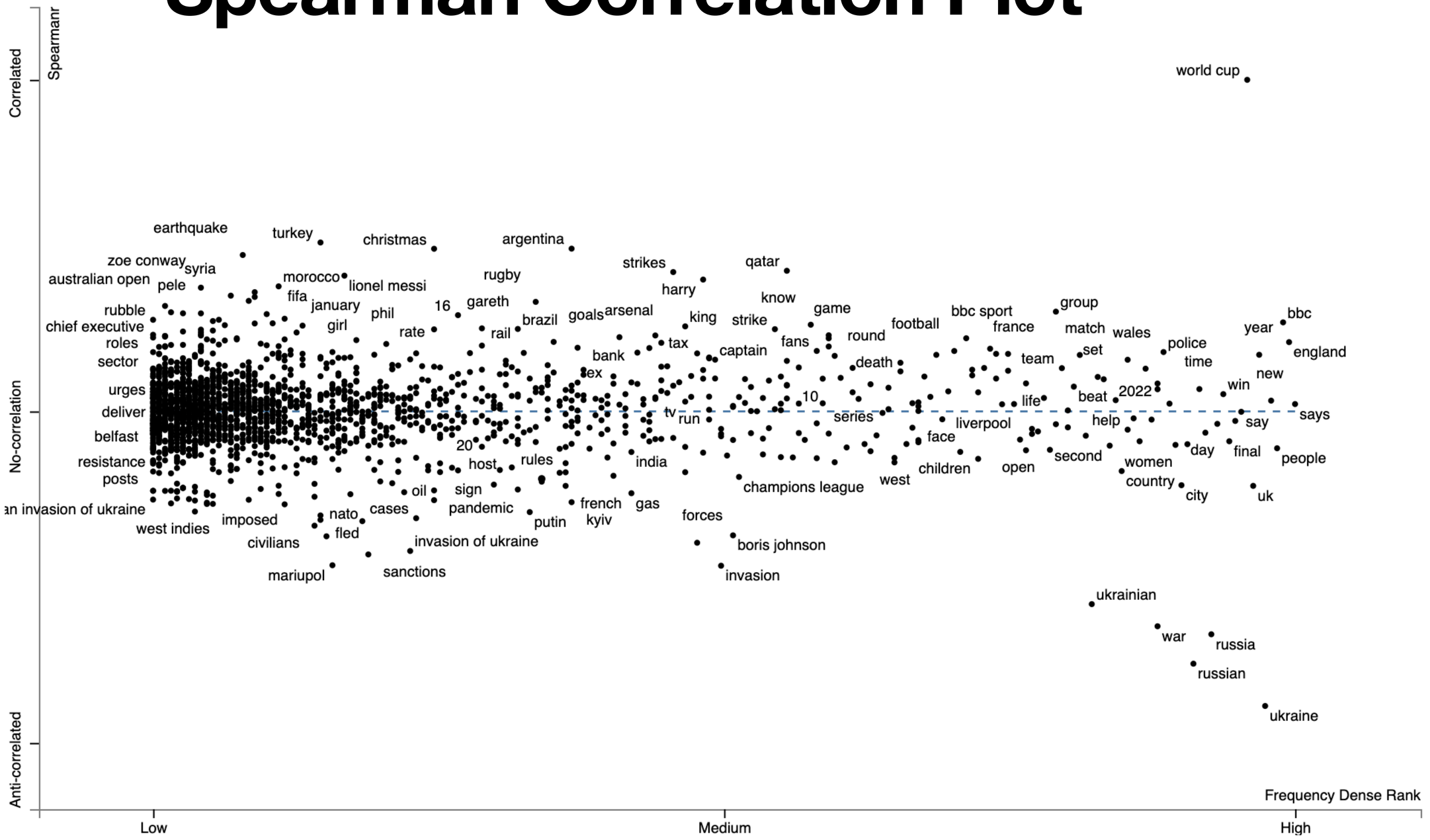
- Position-sensitive dispersion metrics
- Ways of factoring out or calling out periodicity
- Chart interactivity to show a time step-specific plot

Thank you! Any questions?

Phrase Selection

- Take 1-5 grams, keeping those which
  - start and end with a content word
  - appear at least three times
  - keep the top 2000 highest NPMI phrases per phrase length in tokens
  - does not makeup at least 60% of a longer (token-wise) selected phrase
- Select 2000 phrases for visualization by round-robin selection of the highest JSD association for each year
  - Jenson-Shanon Divergence tends to favor topical but high-frequency terms

Spearman Correlation Plot



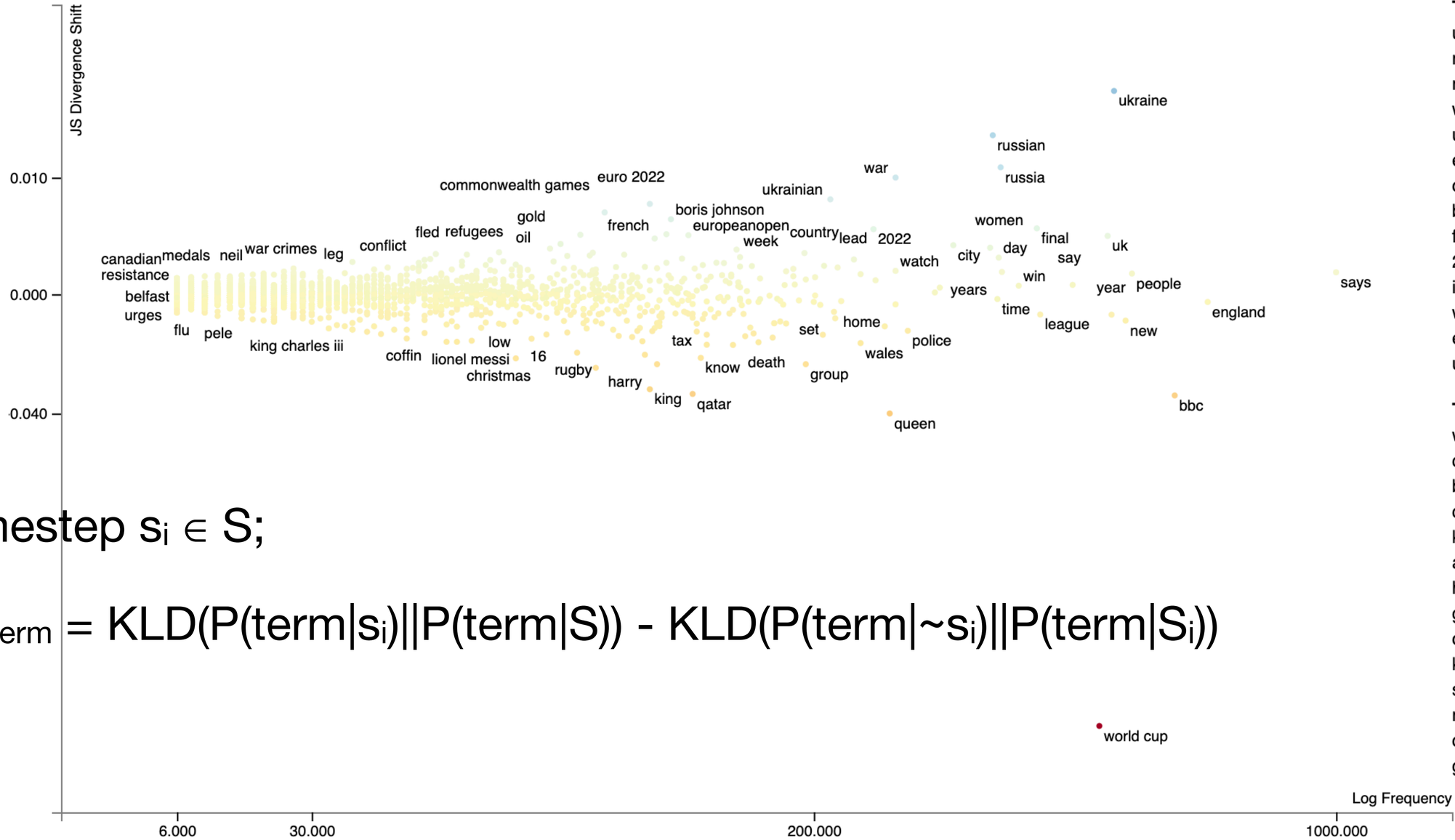
Most correlated

world cup
turkey
argentina
christmas
earthquake
qatar
strikes
lionel messi
harry
morocco
zoe conway
t20 world cup
syria
nurses

Most anti-correlated

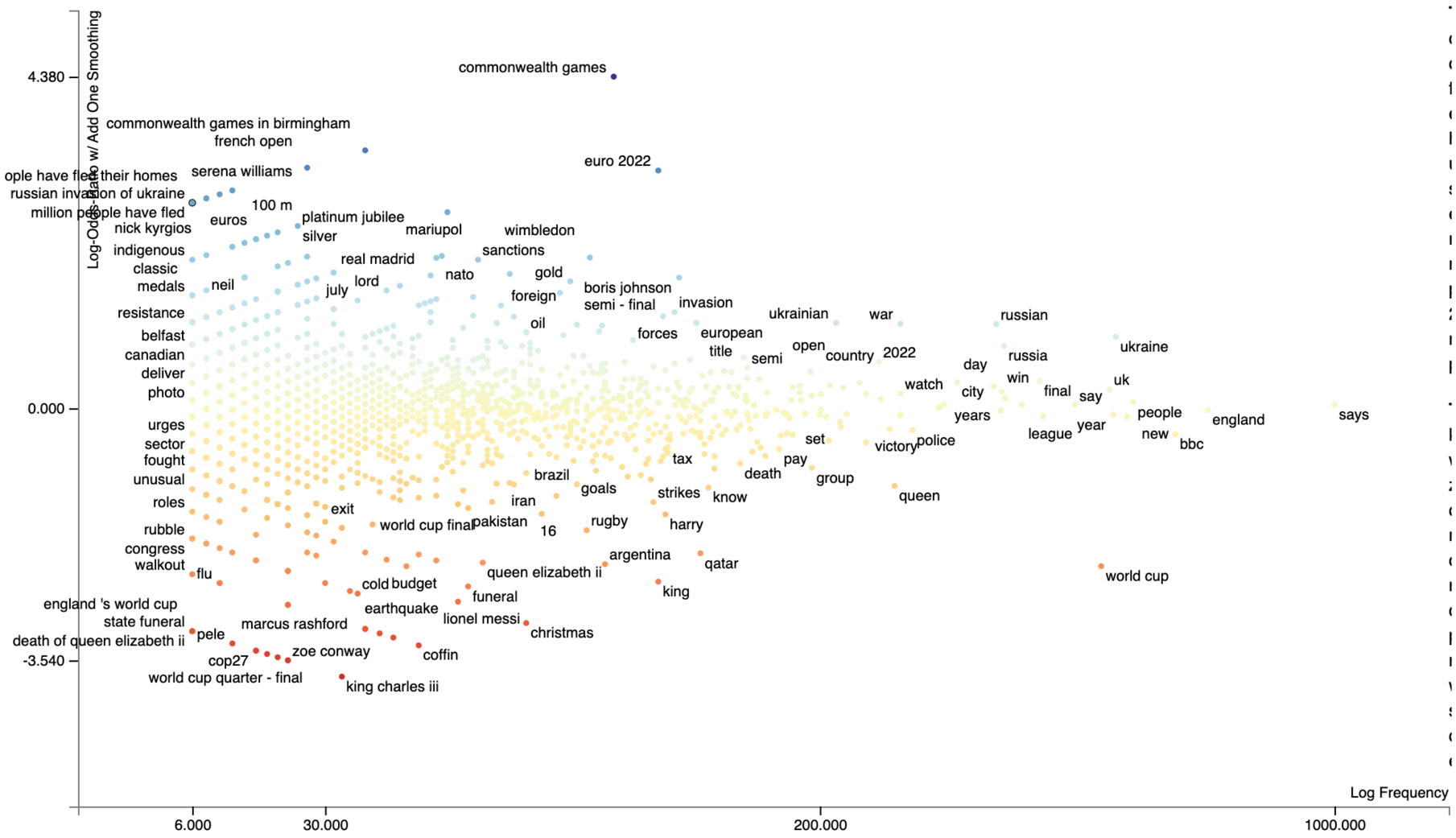
ukraine
russian
russia
war
ukrainian
invasion
mariupol
sanctions
invasion of ukraine
forces
refugees
boris johnson
civilians
russian forces

Delta JS Divergence



- For a timestep $s_i \in S$;
- $\Delta JSD_{term} = KLD(P(term|s_i)||P(term|S)) - KLD(P(term|\sim s_i)||P(term|S_i))$

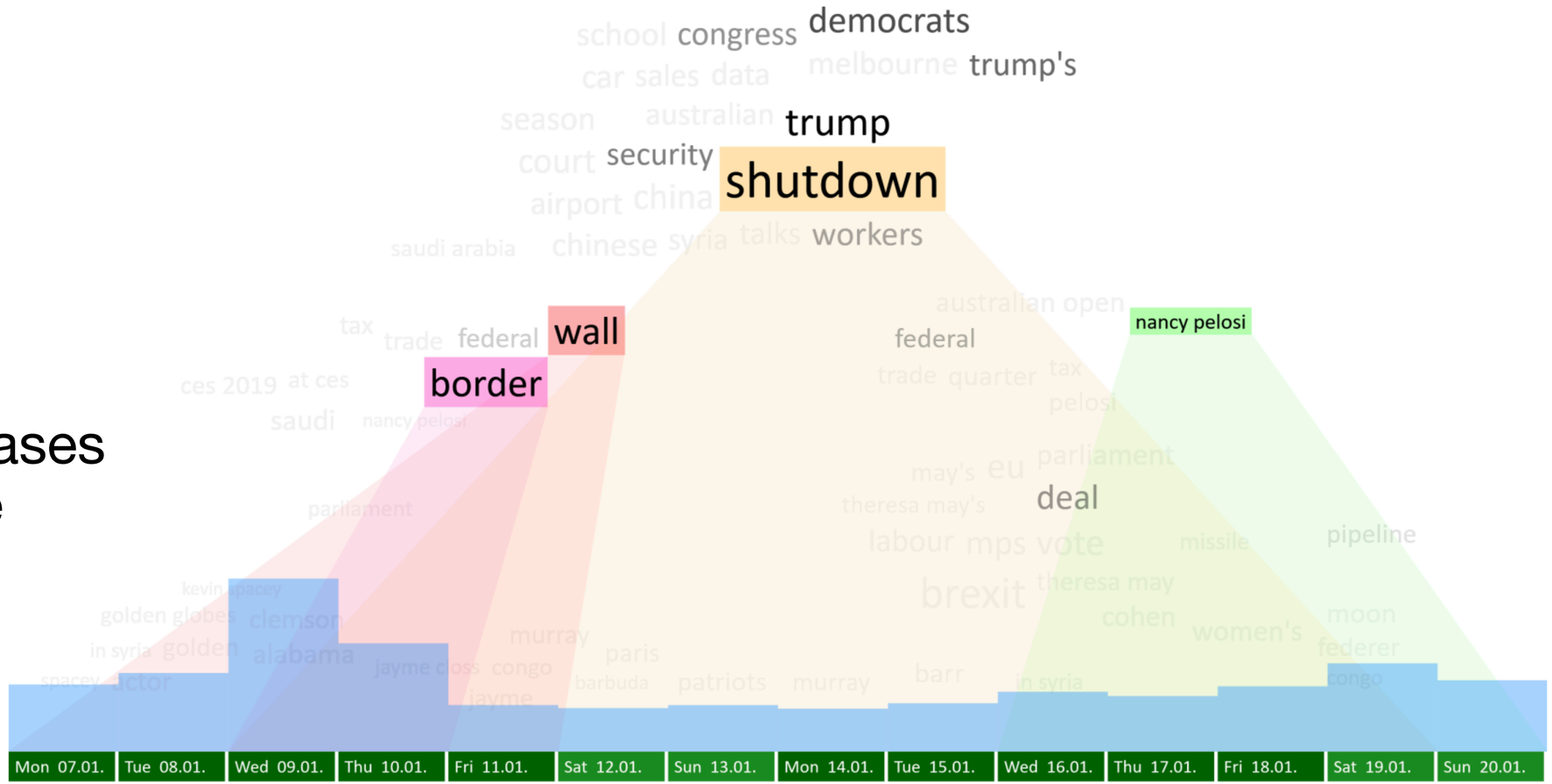
C/f log odds ratio



Related Work

PyramidTags

- Hovering over a phase shows its timespan
- Related phrases are highlighted proportional relatedness
  - Relatedness: two phrases tend to occur in same documents in close proximity



Knittel, J., Koch, S., & Ertl, T. 2021. PyramidTags: Context-, Time- and Word Order-Aware Tag Maps to Explore Large Document Collections. IEEE Transactions on Visualization and Computer Graphics.

Dispersion

- Some terms are highly specific to particular type of movies. Tend to appear clumped in reviews for particular movies.
  - From low to high frequency: movie names, actors, and genres
- Some terms tend to be appear at a roughly consistent rate in different sections of a corpus.
  - Are not specific to a particular type of movie.
  - The words “audience” or “quite” in movie reviews.
  - Function words (we’d think)

Dispersion

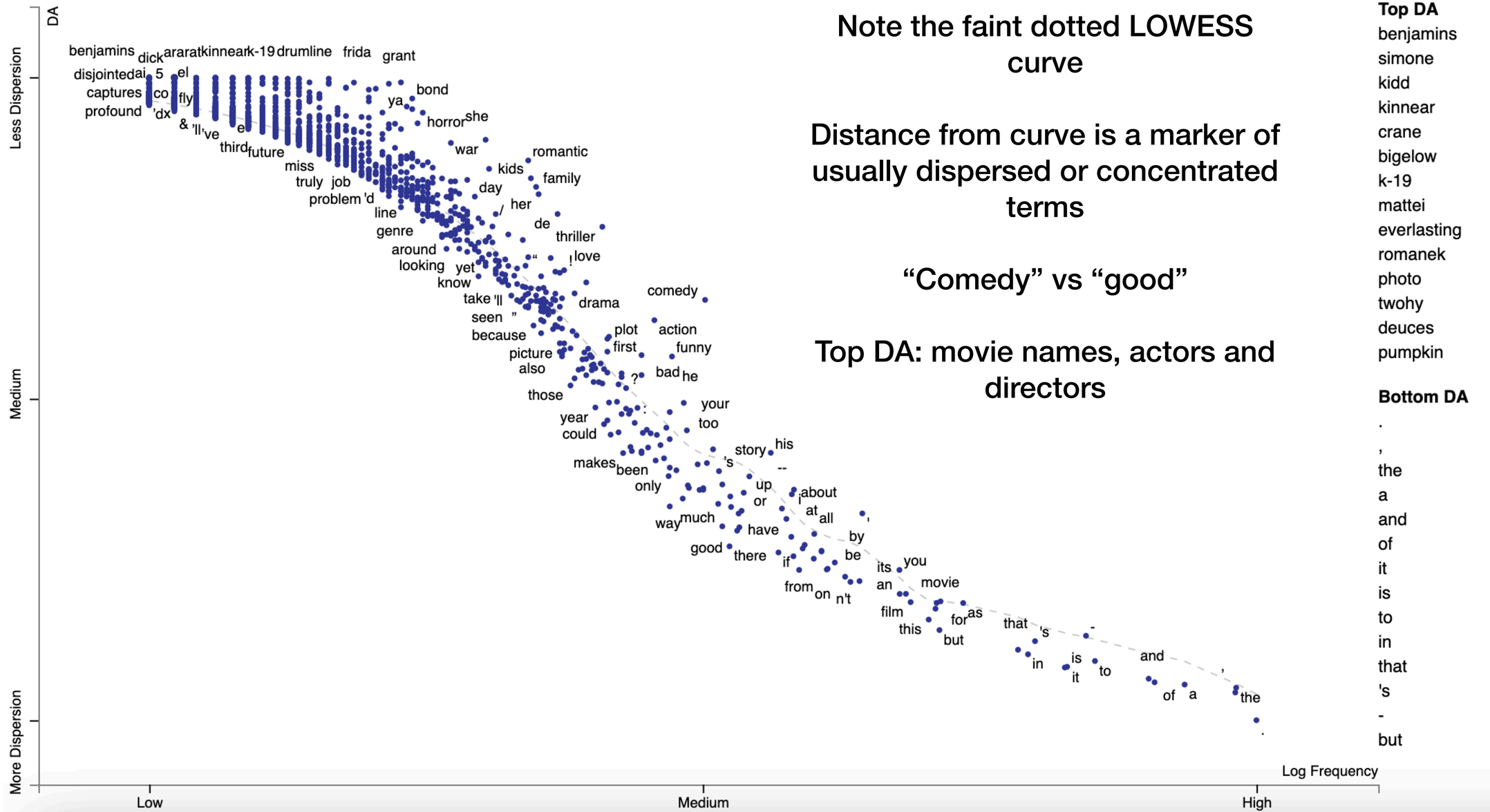
- Measures of dispersion
  - Lots of metrics
- Burch's DA (2017): pairwise difference of absolute deviations.
  - **Assumes roughly equal sized parts**
  - Range: [0, 1] (0: max dispersion)
  - DA = 0.5 when dispersion = thee average term frequency
- Inversely correlated with term frequency
- O(|parts|^2) required to compute

Y_i: term frequency in part i
k: # parts in corpus

$$D_A = 1 - \frac{\frac{1}{k(k-1)/2}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} |Y_i - Y_j|}{2\overline{Y}}$$

Burch, B., Egbert, J., & Biber, D. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*

Dispersion in Movie Corpus



Note the faint dotted LOWESS curve

Distance from curve is a marker of usually dispersed or concentrated terms

“Comedy” vs “good”

Top DA: movie names, actors and directors

Top DA

benjamins
simone
kidd
kinnear
crane
bigelow
k-19
mattei
everlasting
romanek
photo
twohy
deuces
pumpkin

Bottom DA

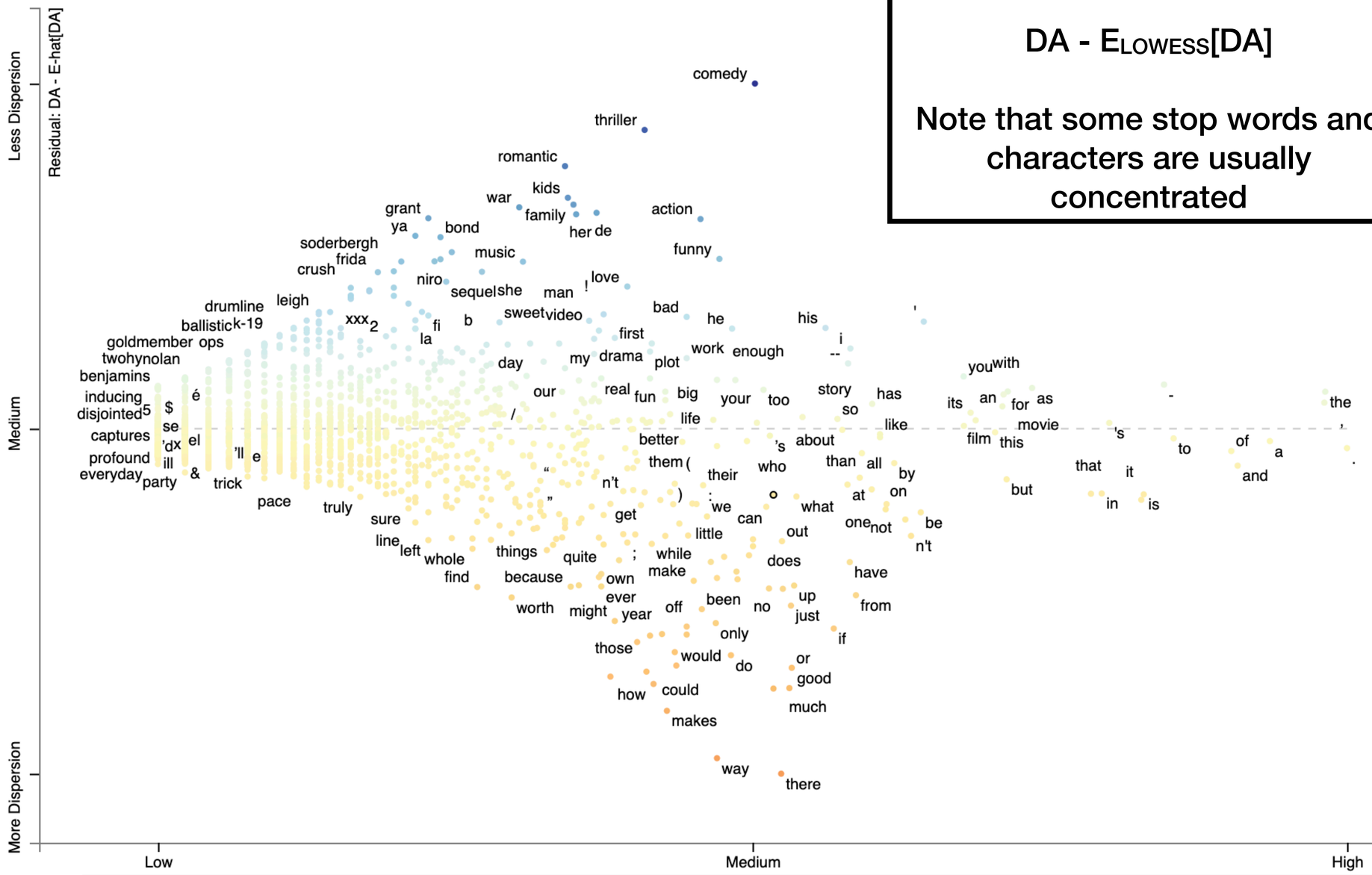
.
,
the
a
and
of
it
is
to
in
that
's
-
but

Residual Dispersion

Y-axis now vertical distance from LOWESS Curve

DA - E_{LOWESS}[DA]

Note that some stop words and characters are usually concentrated



Lower than Expected

comedy
thriller
romantic
kids
family
war
de
her
grant
action
ya
bond
horror
funny

More than Expected

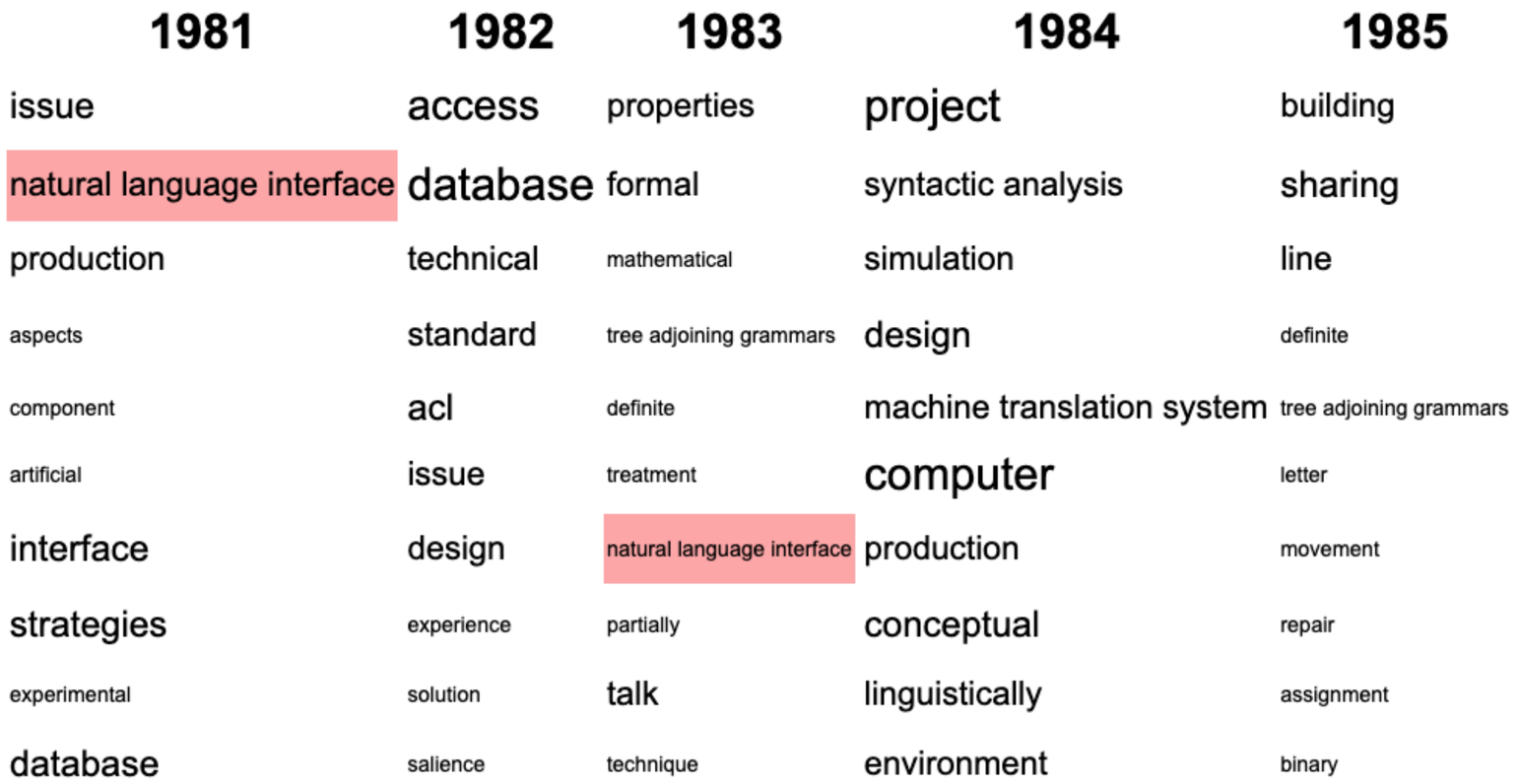
there
way
makes
much
good
how
those
could
or
would
do
may
year
off

ACL Paper Title Corpus

- 10,974 long ACL/EMNLP paper titles from 1979 to 2022
- Objective: see changes in the field of NLP
- Consider year published as category

Visualizing Over Time

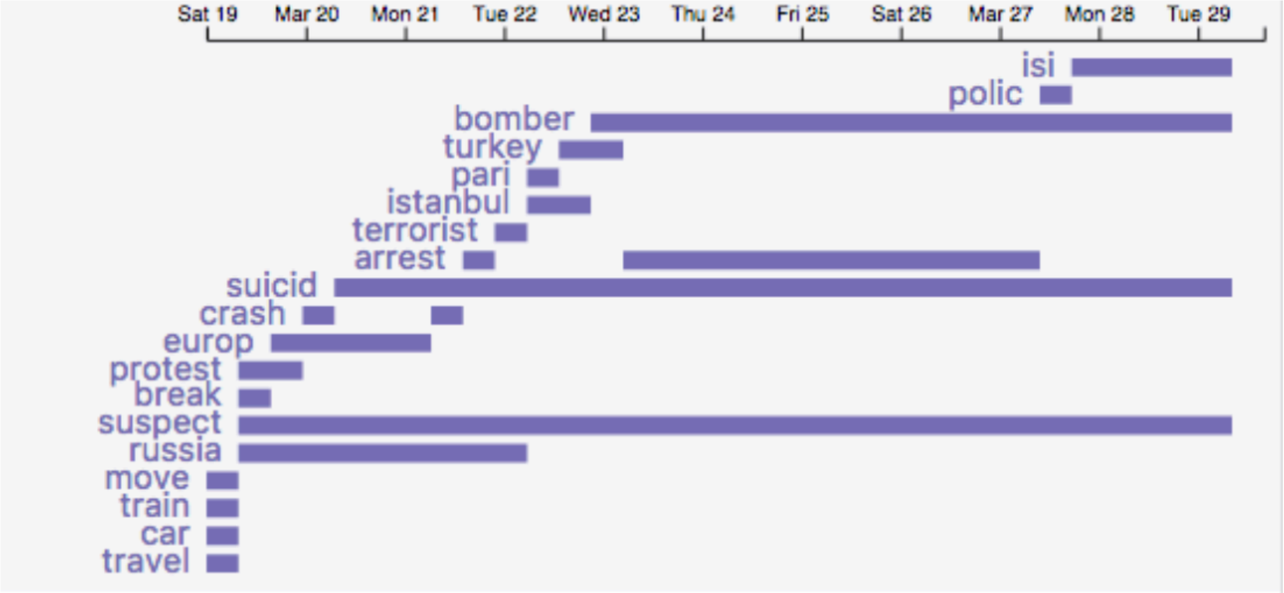
- We can make visualization similar to parallel tag clouds
- Ordered by keyness (log-odds-ratio)
- Size frequency in category (globally scaled)
- Too many years to view without scrolling



Related Work

ESTEEM

- Uses a Gantt chart to display words with similar embeddings to a query
- Embeddings are produced at each time step
- Simple, easy-to-follow visualization
- Few terms can be visualized this way



- Stemmed words with similar day-specific embeddings to “bomb” from German Tweets after the 2016 Brussels terrorist attack

Dustin Arendt and Svitlana Volkova. 2017. ESTEEM: A Novel Framework for Qualitatively Evaluating and Visualizing Spatiotemporal Embeddings in Social Media. In Proceedings of ACL 2017, System Demonstrations.

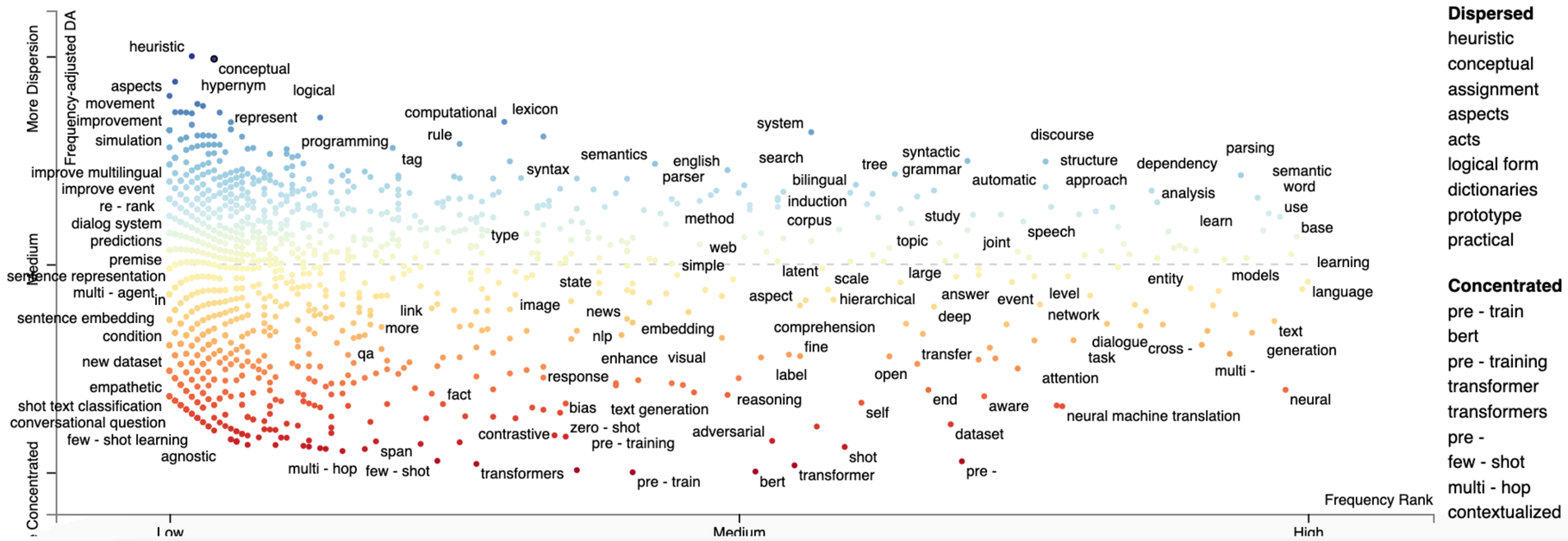
Visualizing Over Time

- Solve scrolling problem: merge years together with high cosine similarity of keyness values of displayed terms

1979-1983	1984-1988	1989-2006	2007-2008	2009-2010	2011-2015	2016-2018	2019-2022
natural language interface	project	unification - base	topic segmentation	web search	joint inference	attention model	bert
design	tree adjoining grammars	maximum entropy	phoneme conversion	expectation	distributional semantics	lstm	pre - train
issue	machine translation system	unification	semantic information	letter	segmentation use	neural semantic	pre - training
access	computer	automatic acquisition	online learning	phrase - base statistical machine	dual decomposition	deep reinforcement learning	transformers
production	sharing	lr	non - local	system combination	decipherment	sequence learning	few - shot
database	design	utterances	semantic relatedness	hierarchical phrase - base translation	acl	skip	multi - hop
computer	natural language interface	processor	context - sensitive	binarization	reorder	sequence models	pretraine
speech act	production	corpus - base	em	base statistical machine translation	recursive neural	deep neural networks	pretrained
interface	unification grammar	machine learning	improve statistical machine translation	phrase - base translation	unsupervised semantic	bandit	pre - train language model
experimental	categorial	constraint - base	subjectivity	semi - supervised learning	walk	character - level	contrastive learning

Use residual dispersion to see interesting terms which do not make the tag clouds

1979-1983	1984-1988	1989-2008	2009-2010	2011-2015	2016-2018	2019-2022
natural language interface	project	machine learning	web search	joint inference	attention model	bert
design	tree adjoining grammars	unification - base	expectation	distributional semantics	lstm	pre - train
issue	machine translation system	automatic acquisition	letter	segmentation use	neural semantic	pre - training
access	computer	lr	phrase - base statistical machine	dual decomposition	deep reinforcement learning	transformers
production	sharing	maximum entropy	system combination	decipherment	sequence learning	few - shot
database	design	lexicalized	hierarchical phrase - base translation	acl	skip	multi - hop
computer	natural language interface	corpus - base	binarization	reorder	sequence models	pretraine
speech act	production	large corpora	base statistical machine translation	recursive neural	deep neural networks	pretrained
interface	unification grammar	co - occurrence	phrase - base translation	unsupervised semantic	bandit	pre - train language model
experimental	categorial	unification	semi - supervised learning	walk	character - level	contrastive learning



Dispersed
heuristic
conceptual
assignment
aspects
acts
logical form
dictionaries
prototype
practical

Concentrated
pre - train
bert
pre - training
transformer
transformers
pre -
few - shot
multi - hop
contextualized