

ICAME 44

English going places, corpora crossing spaces



Understanding corpus text prototypicality

A multifaceted problem

Laurence Anthony (Waseda University)
anthony@waseda.jp

Nick Smith (University of Leicester)
ns359@leicester.ac.uk

Sebastian Hoffmann (Universität Trier)
hoffmann@uni-trier.de

Paul Rayson (Lancaster University)
p.rayson@lancaster.ac.uk



May 18 (Thu), 2023. Hybrid
<https://humanities.nwu.ac.za/languages/ICAME44>

Overview

- Background
 - definitions and characteristics of prototypicality
 - importance of prototypicality
 - Identification of prototypical texts
- RQs and experiment design
 - Replicating the Anthony & Baker (2015) study
 - Expanding the study across multiple annotation layers
- Results and discussion
 - Prototypical short/long texts
 - Outlier texts in a 1 m word corpora
 - Improving the method for more nuanced rankings
- Conclusions

ProtAnt
A tool for analysing the prototypicality of texts

Laurence Anthony and Paul Baker
Waseda University / Lancaster University

The screenshot shows the ProtAnt application window. On the left, there's a list of text items with columns for 'File', 'Key Type', 'Key Score', 'Non-Key Type', 'Non-Key Score', 'All Type', and 'All Score'. The main area displays a detailed view of a selected item, showing its text and various annotations. At the bottom, there are controls for 'Review', 'Rank', and 'Export'.

Outlier File	Ranking
K12	40/100
L9	29/30
P13	18/18
C8	18/18
N7	18/17
A3	29/30
M6	40/76
L13	30/31
R8	8/50
L15	25/30
C6	26/26
N8	4/7
A7	29/30
A5	29/30
L2	18/18
C8	18/18
N7	18/17
A3	29/30
M6	40/76
L13	30/31
R8	8/50
L15	25/30
C6	26/26
N8	4/7
A7	29/30
A5	29/30
L2	18/18

Background

definitions and characteristics of prototypicality;
importance of prototypicality; Identification of prototypical texts

The screenshot shows the ProAnt 1.2.0 software interface. On the left, there is a list of files being analyzed, including 1.txt through 15.txt. Below this is a 'Reference Corpus' section with a list of files from A01.txt to A09.txt. The main area on the right displays a table of results for each file, with columns for File, KeyTypes, KeyItems, NormKeyTypes, NormKeyItems, AllTypes, and AllItems. Below the table is a 'Keywords Statistics' section with a table of keywords, their frequencies, and effects.

File	KeyTypes	KeyItems	NormKeyTypes	NormKeyItems	AllTypes	AllItems
1.txt	22	67	47.2973	843.762	444	1544
2.txt	20	62	46.9484	62	426	1000
3.txt	20	45	41.3223	38.4287	484	1171
4.txt	22	90	41.1983	74.5033	234	1228
5.txt	17	132	37.2807	111.864	456	1180
6.txt	18	58	33.1492	47.0153	543	1213
7.txt	14	40	27.3138	41.9929	513	866

Keyword	Freq	KeyItems	Effect
betan	12	130.038	151.310
betahat	10	101.088	1194.74
jeffrey	12	91.7888	1075.34
hufe	9	68.8404	806.391
farah	9	68.8404	806.391
hurdhani	9	68.8404	806.391
setra	7	53.542	627.103



Background

Definitions and facets of prototypicality

"having the **typical qualities** of a particular group or kind of person or thing"
(Merriam-Webster, 2014)

the **clearest, best, most typical, most representative** examples
(Labov 1973, Rosch 1975, Gries 2001)

■ Prototypicality as "graded centrality"

[Croft, W., & Cruse, D. A. (2004). Cognitive linguistics. Cambridge University Press]

- "Members that are judged to be the best examples of a category can be considered to be the most central in the category" (p. 77)
 - e.g., Goodness-of-Exemplar (GOE) rankings for VEGETABLE
 - leek, carrot (GOE rating 1)
 - lemon (GOE rating 7)
- Correlates with frequency and order of mention, order of learning, family resemblance, verification speed, priming

Background

Definitions and facets of prototypicality

- Facets of prototypicality

[Croft et al. 2004; Lakoff 1987: 84-90]

- Stereotypicality
 - "the shape of a diamond"
- Closeness to an ideal
 - "the perfect diamond shape"
- Typicality/representativeness
 - "the most common diamond shape"



The Hope Diamond



Round



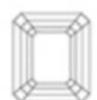
Pear



Oval



Cushion



Emerald



Radiant



Asscher



Princess



Heart



Marquise

Background

Definitions and facets of prototypicality - in corpus linguistics

- Types of prototypicality
 - Stereotypicality; Closeness to an ideal; Typicality/representativeness
- Properties of language
 - lexical, grammatical, structural, semantic, contextual, functional, thematic, ...
 - lexical – single words vs multi-word units



Background

Importance of prototypicality

- corpus creation
 - choosing appropriate texts to include in a corpus
 - identifying problematic texts to exclude from a corpus
- corpus analysis
 - choosing texts for close reading (down-sampling) to...
 - formulate hypotheses
 - validate findings created at the corpus level
- pedagogic purposes
 - selecting 'good' examples of texts to serve as in-class models
 - selecting atypical/outlier learner texts to identify language problems
- ...

Background

Identification of prototypical texts

- Critical Discourse Analysis (CDA) and other qualitative studies
 - opportunistic selection
 - e.g., Caldas-Coulthard et al. (2003)
 - "...we purchased all the 15 bear books available in a local children's bookstore in London."
 - limitations
 - non-principled
 - possible bias of researcher ('cherry picking')
 - difficult to replicate the results

Background

Identification of prototypical texts

- Critical Discourse Analysis (CDA) and other qualitative studies
 - selective downsizing
 - e.g. Khosravini (2010)
 - in a corpus of 170,000 articles, select articles from five one-week periods where the number of articles about immigration peak (resulting in 439 articles)
 - limitations
 - can still result in a large number of sample texts
 - 'cherry picking' criticism is not completely addressed

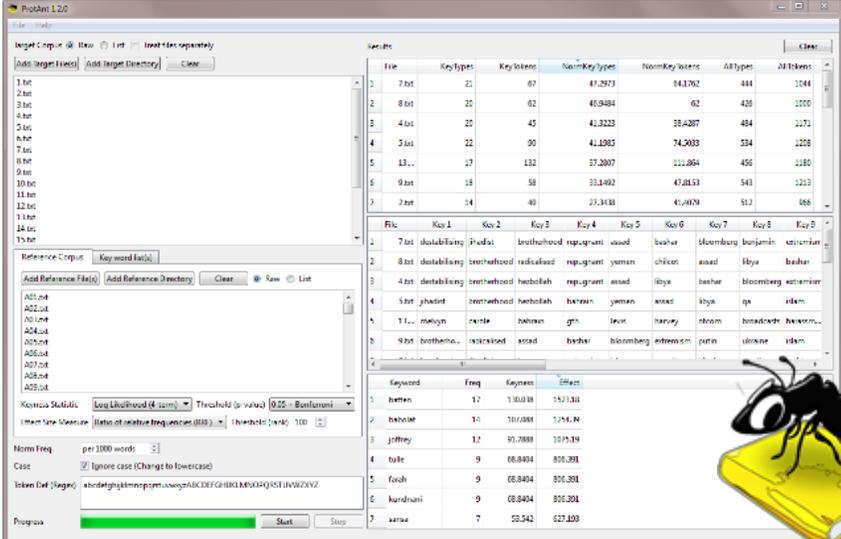
Background:

Identification of prototypical texts

- *ProtAnt* (Anthony & Baker, 2015)

<http://www.laurenceanthony.net/software/antcorgen/>

- a freeware automatic text prototype detection tool
 - ranks corpus texts by degree of 'lexical' prototypicality
 - e.g., number of (normed) keywords per text
 - e.g., number of pre-defined 'key' words per text (e.g., AWL list)
 - displays keyword lists, per-text lexical profiles, ranking criteria
 - allows for easy close reading of ranked texts



The screenshot shows the ProtAnt 1.2.0 software interface. The main window displays a list of ranked texts with columns for File, Key Types, Key Items, NormKey Types, NormKey Items, All Types, and All Items. Below this, a detailed lexical profile is shown for a selected text, including a list of keywords and their frequencies. The interface also includes a Reference Corpus section and various settings for analysis.

File	Key Types	Key Items	NormKey Types	NormKey Items	All Types	All Items
1.txt	21	67	47,2973	84,1762	444	1544
2.txt	20	62	46,9484	62	426	1020
3.txt	20	40	41,3223	38,4287	484	1171
4.txt	22	90	41,1883	74,2033	234	1228
5.txt	17	132	37,2807	111,864	456	1180
6.txt	18	58	10,9402	47,8133	543	1213
7.txt	14	40	27,3438	41,4070	513	886

File	Key 1	Key 2	Key 3	Key 4	Key 5	Key 6	Key 7	Key 8	Key 9
1.txt	detabliang	Pudisi	brotherhood	repugnant	asad	barbar	Bloomberg	benjamin	strumian
2.txt	detabliang	brotherhood	radicalised	repugnant	syman	chikus	asad	ibya	barbar
3.txt	detabliang	brotherhood	herbolah	repugnant	asad	ibya	barbar	Bloomberg	strumian
4.txt	ghadid	brotherhood	herbolah	barazan	syman	asad	ibya	qs	idam
5.txt	malign	rande	barazan	gts	ibya	harvey	strom	broadside	barazan
6.txt	brotherho	radicalised	asad	barbar	Bloomberg	edwison	putin	ultraone	idam

Keyword	Freq	Keyness	Effect
barazan	17	1/8138	1571.516
broadside	10	10/1089	1744.84
joffrey	12	91.7888	1075.19
ibya	9	68.8404	806.301
foresh	9	68.8404	806.301
handshani	9	68.8404	806.301
syman	7	55.542	627.100

ProtAnt (Anthony & Baker, 2015)

Basic algorithm

- **Step 1:** Generate **keywords** for the **target corpus**
 - e.g., using log-likelihood + (log) relative frequency (effect size) against a reference corpus
 - e.g., using a pre-defined list of 'key' words (GSL 1/2, AWL, ...)
- **Step 2:** Rank **target files** by the (normalized) number of **keywords** they contain
 - e.g., (key types in file)/(total types in file)
 - e.g., (key tokens in file)/ (total tokens in file)
 - e.g., (log key types | tokens)/(log total types | tokens)
- **Step 3:** Display profiling information to the user
 - the keyword list
 - the per-file keyword list
 - the target file rankings

The screenshot displays the ProtAnt 1.2.0 application window. The main window is titled 'ProtAnt 1.2.0' and contains several panels. On the left, there is a 'Target Corpus' list with files like 1.txt, 2.txt, etc. Below it is a 'Reference Corpus' list with files like A01.txt, A02.txt, etc. The bottom left panel shows 'Norm Freq' settings, including 'per 1000 words' and 'Ignore case (Change to lowercase)'. The bottom right panel shows 'Token Def (Regex)' with a complex regular expression. The main area on the right is a table showing the results of the analysis, including file rankings and keyword lists.

File	Key Types	Key Tokens	NormKey Types	NormKey Tokens	All Types	All Tokens
1.txt	21	67	47.2973	843.762	444	1544
2.txt	20	62	46.9484	62	426	1020
3.txt	22	40	41.3223	38.4287	484	1171
4.txt	20	90	41.1883	74.2033	234	1228
5.txt	17	132	37.2807	111.864	456	1180
6.txt	18	58	33.8402	47.8133	543	1213
7.txt	14	40	27.3438	41.4070	513	886

File	Key 1	Key 2	Key 3	Key 4	Key 5	Key 6	Key 7	Key 8	Key 9
1.txt	detab/fang	P-ed/ai	broth/hood	repugnant	assad	barbar	Bloomberg	benjamin	strawman
2.txt	detab/fang	broth/hood	radical/ed	repugnant	syman	chicot	assad	ibya	barbar
3.txt	detab/fang	broth/hood	herbolah	repugnant	assad	ibya	barbar	Bloomberg	strawman
4.txt	ghed/it	broth/hood	herbolah	barzain	syman	assad	ibya	qs	idam
5.txt	malign	rande	barzain	grt	ibya	harvey	strom	broadside	barzain...
6.txt	broth/hood	radical/ed	assad	barbar	Bloomberg	edwison	putin	ultra/na	idam

Keyword	Freq	Key/len	Effect
barzain	17	1/8/10/18	1571.516
broadside	10	10/1/8/8	1714.14
ghed/it	12	9/1/8/8	1075.19
ibya	9	08/8/04	806.391
qs	9	08/8/04	806.391
strawman	9	08/8/04	806.391
ultra/na	7	05/5/42	627.100

Research Questions

Research Questions

1. What can a replication and extension of the Anthony & Baker (2015) study tell us about prototypicality?
What impact do different layers of annotation have on text rankings?
 - lexical (LEX) items
 - lemma (LEM) items
 - USAS semantic (SEM) tags
 - CLAWS 7 part-of-speech (POS) tags
2. How can *ProtAnt* be improved to allow more nuanced rankings?
3. What are the implications of corpus prototypicality for corpus design and methods?

RQ1

What can a replication and extension of the Anthony & Baker (2015) study tell us about prototypicality?
What impact do different layers of annotation have on text rankings?

Experiment 1: Prototypicality rankings of 'in/out' texts

'Islam' news article corpus - LEX rankings

row_id	p05	p01	p001	p0001
1	04_islam.txt	07_islam.txt	08_islam.txt	08_islam.txt
2	05_islam.txt	08_islam.txt	07_islam.txt	07_islam.txt
3	08_islam.txt	05_islam.txt	04_islam.txt	04_islam.txt
4	07_islam.txt	04_islam.txt	05_islam.txt	05_islam.txt
5	15_football.txt	06_islam.txt	06_islam.txt	06_islam.txt
6	09_islam.txt	15_football.txt	03_islam.txt	13_football.txt
7	02_islam.txt	09_islam.txt	13_football.txt	03_islam.txt
8	06_islam.txt	03_islam.txt	09_islam.txt	09_islam.txt
9	03_islam.txt	02_islam.txt	11_football.txt	10_islam.txt
10	01_islam.txt	10_islam.txt	02_islam.txt	11_football.txt
11	13_football.txt	13_football.txt	10_islam.txt	18_tennis.txt
12	12_football.txt	01_islam.txt	18_tennis.txt	02_islam.txt
13	19_review.txt	19_review.txt	15_football.txt	01_islam.txt
14	10_islam.txt	14_football.txt	01_islam.txt	15_football.txt
15	20_art.txt	12_football.txt	12_football.txt	12_football.txt
16	16_obituary.txt	20_art.txt	20_art.txt	16_obituary.txt
17	18_tennis.txt	11_football.txt	14_football.txt	20_art.txt
18	11_football.txt	18_tennis.txt	16_obituary.txt	14_football.txt
19	14_football.txt	16_obituary.txt	19_review.txt	19_review.txt
20	17_science.txt	17_science.txt	17_science.txt	17_science.txt

Log Likelihood (LL2)

- Results confirm accurate rankings of 'in' texts for a pseudo corpus focused on the topic of 'Islam'
- The 5 texts ranked as most prototypical ($p < .0001$) report or comment on a speech about 'radical Islam' by UK prime minister Tony Blair
- Why are texts about 'football' ranked so high? [Keywords about 'football' 'pollute' the results?]

Experiment 1: Prototypicality rankings of 'in/out' texts

'Islam' news article corpus - LEM rankings

row_id	p05	p01	p001	p0001
1	05_islam.txt	07_islam.txt	07_islam.txt	07_islam.txt
2	15_football.txt	08_islam.txt	08_islam.txt	08_islam.txt
3	04_islam.txt	05_islam.txt	04_islam.txt	05_islam.txt
4	08_islam.txt	04_islam.txt	05_islam.txt	04_islam.txt
5	07_islam.txt	15_football.txt	06_islam.txt	09_islam.txt
6	09_islam.txt	06_islam.txt	09_islam.txt	13_football.txt
7	06_islam.txt	09_islam.txt	13_football.txt	03_islam.txt
8	02_islam.txt	02_islam.txt	03_islam.txt	06_islam.txt
9	03_islam.txt	10_islam.txt	02_islam.txt	02_islam.txt
10	10_islam.txt	13_football.txt	10_islam.txt	11_football.txt
11	01_islam.txt	03_islam.txt	11_football.txt	10_islam.txt
12	19_review.txt	19_review.txt	15_football.txt	18_tennis.txt
13	13_football.txt	01_islam.txt	01_islam.txt	15_football.txt
14	12_football.txt	11_football.txt	18_tennis.txt	01_islam.txt
15	20_art.txt	12_football.txt	12_football.txt	14_football.txt
16	16_obituary.txt	20_art.txt	20_art.txt	12_football.txt
17	18_tennis.txt	14_football.txt	14_football.txt	20_art.txt
18	11_football.txt	16_obituary.txt	16_obituary.txt	16_obituary.txt
19	14_football.txt	18_tennis.txt	19_review.txt	19_review.txt
20	17_science.txt	17_science.txt	17_science.txt	17_science.txt

Log Likelihood (LL2)

- Results confirm accurate rankings of 'in' texts for a pseudo corpus focused on the topic of 'Islam'
- LEM rankings differ slightly for LEX rankings
- LEM rankings ($p < .05$) **improve** on LEX rankings
- Why are texts about 'football' ranked so high? [Keywords about 'football' 'pollute' the results?]

Experiment 1: Prototypicality rankings of 'in/out' texts

'Islam' news article corpus - SEM rankings

row_id	p05	p01	p001	p0001
1	06_islam.txt	03_islam.txt	03_islam.txt	06_islam.txt
2	03_islam.txt	06_islam.txt	06_islam.txt	03_islam.txt
3	10_islam.txt	10_islam.txt	13_football.txt	14_football.txt
4	05_islam.txt	02_islam.txt	15_football.txt	13_football.txt
5	02_islam.txt	05_islam.txt	10_islam.txt	11_football.txt
6	14_football.txt	14_football.txt	14_football.txt	15_football.txt
7	04_islam.txt	01_islam.txt	05_islam.txt	02_islam.txt
8	15_football.txt	08_islam.txt	02_islam.txt	10_islam.txt
9	07_islam.txt	04_islam.txt	11_football.txt	18_tennis.txt
10	11_football.txt	15_football.txt	08_islam.txt	08_islam.txt
11	09_islam.txt	13_football.txt	01_islam.txt	01_islam.txt
12	13_football.txt	07_islam.txt	04_islam.txt	05_islam.txt
13	08_islam.txt	11_football.txt	18_tennis.txt	12_football.txt
14	01_islam.txt	19_review.txt	12_football.txt	09_islam.txt
15	19_review.txt	09_islam.txt	09_islam.txt	20_art.txt
16	18_tennis.txt	18_tennis.txt	19_review.txt	04_islam.txt
17	17_science.txt	12_football.txt	07_islam.txt	07_islam.txt
18	16_obituary.txt	17_science.txt	17_science.txt	19_review.txt
19	12_football.txt	16_obituary.txt	20_art.txt	16_obituary.txt
20	20_art.txt	20_art.txt	16_obituary.txt	17_science.txt

Log Likelihood (LL2)

- Results confirm accurate rankings of 'in' texts for a pseudo corpus focused on the topic of 'Islam' (as the p value is increased)
- Fewer SEM key items lead to 'unstable' rankings
- Why are texts about 'football' ranked so high? [Keywords about 'football' 'pollute' the results?]

Experiment 1: Prototypicality rankings of 'in/out' texts

'Islam' news article corpus - POS rankings

row_id	p05	p01	p001	p0001
1	08_islam.txt	08_islam.txt	08_islam.txt	08_islam.txt
2	10_islam.txt	07_islam.txt	07_islam.txt	10_islam.txt
3	06_islam.txt	18_tennis.txt	16_obituary.txt	07_islam.txt
4	02_islam.txt	10_islam.txt	18_tennis.txt	13_football.txt
5	18_tennis.txt	06_islam.txt	10_islam.txt	05_islam.txt
6	04_islam.txt	02_islam.txt	06_islam.txt	09_islam.txt
7	05_islam.txt	09_islam.txt	13_football.txt	16_obituary.txt
8	14_football.txt	04_islam.txt	02_islam.txt	18_tennis.txt
9	07_islam.txt	05_islam.txt	05_islam.txt	02_islam.txt
10	16_obituary.txt	16_obituary.txt	09_islam.txt	06_islam.txt
11	03_islam.txt	13_football.txt	04_islam.txt	20_art.txt
12	09_islam.txt	01_islam.txt	03_islam.txt	19_review.txt
13	13_football.txt	14_football.txt	20_art.txt	04_islam.txt
14	01_islam.txt	03_islam.txt	19_review.txt	01_islam.txt
15	17_science.txt	15_football.txt	01_islam.txt	14_football.txt
16	11_football.txt	17_science.txt	14_football.txt	12_football.txt
17	20_art.txt	11_football.txt	12_football.txt	11_football.txt
18	12_football.txt	20_art.txt	17_science.txt	15_football.txt
19	19_review.txt	19_review.txt	11_football.txt	03_islam.txt
20	15_football.txt	12_football.txt	15_football.txt	17_science.txt

Log Likelihood (LL2)

- Results confirm accurate rankings of 'in' texts for a pseudo corpus focused on the topic of 'Islam' (as the p value is increased)
- Fewer POS key items leads to 'unstable' rankings
- Why are texts about 'football' ranked so high? [Keywords about 'football' 'pollute' the results?]

Experiment 2: Ranking of 'outlier' texts

AmE06 categories + 1 outlier category text (LEX)

Category	Register	Outlier File	Ranking	Total Files	Diff
A	Press: Reportage	K12	45	45	0
B	Press: Editorial	L9	28	28	0
C	Press: Reviews	P13	18	18	0
D	Religion	C8	18	18	0
E	Skills, Trades and Hobbies	N7	36	37	1
F	Popular Lore	A3	19	49	30
G	Belles Lettres, Biographies, Essays	M6	48	76	28
H	Miscellaneous: Government documents, industrial reports etc	L13	31	31	0
J	Academic prose in various disciplines	R8	81	81	0
K	General Fiction	E15	30	30	0
L	Mystery and Detective Fiction	C6	25	25	0
M	Science Fiction	N8	4	7	3
N	Adventure and Western	A7	30	30	0
P	Romance and Love story	A5	30	30	0
R	Humour	L2	2	10	8

Log Likelihood (LL2); $p < 0.0001$

Experiment 2: Ranking of 'outlier' texts

AmE06 categories + 1 outlier category text (LEM)

Category	Register	Outlier File	Ranking	Total Files	Diff
A	Press: Reportage	K12	45	45	0
B	Press: Editorial	L9	28	28	0
C	Press: Reviews	P13	18	18	0
D	Religion	C8	18	18	0
E	Skills, Trades and Hobbies	N7	33	37	4
F	Popular Lore	A3	12	49	37
G	Belles Lettres, Biographies, Essays	M6	62	76	14
H	Miscellaneous: Government documents, industrial reports etc	L13	31	31	0
J	Academic prose in various disciplines	R8	81	81	0
K	General Fiction	E15	30	30	0
L	Mystery and Detective Fiction	C6	25	25	0
M	Science Fiction	N8	5	7	2
N	Adventure and Western	A7	30	30	0
P	Romance and Love story	A5	30	30	0
R	Humour	L2	2	10	8

Log Likelihood (LL2); $p < 0.0001$

Experiment 2: Ranking of 'outlier' texts

AmE06 categories + 1 outlier category text (SEM)

Category	Register	Outlier File	Ranking	Total Files	Diff
A	Press: Reportage	K12	45	45	0
B	Press: Editorial	L9	28	28	0
C	Press: Reviews	P13	18	18	0
D	Religion	C8	14	18	4
E	Skills, Trades and Hobbies	N7	36	37	1
F	Popular Lore	A3	46	49	3
G	Belles Lettres, Biographies, Essays	M6	63	76	13
H	Miscellaneous: Government documents, industrial reports etc	L13	31	31	0
J	Academic prose in various disciplines	R8	81	81	0
K	General Fiction	E15	30	30	0
L	Mystery and Detective Fiction	C6	18	25	7
M	Science Fiction	N8	6	7	1
N	Adventure and Western	A7	30	30	0
P	Romance and Love story	A5	30	30	0
R	Humour	L2	3	10	7

Log Likelihood (LL2); $p < 0.05$

Experiment 2: Ranking of 'outlier' texts

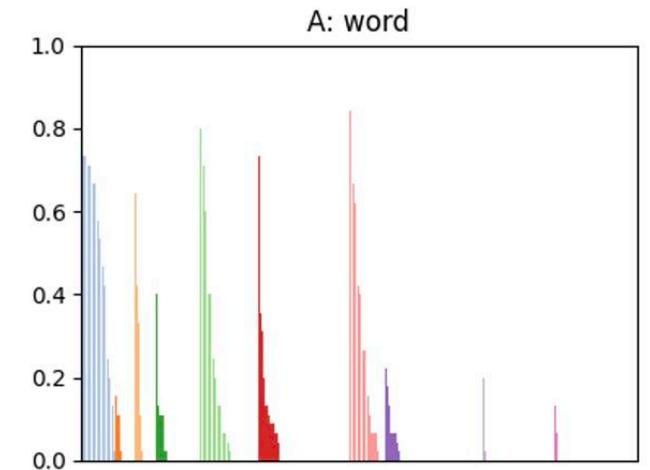
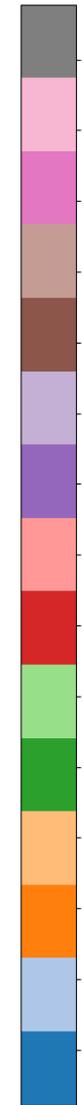
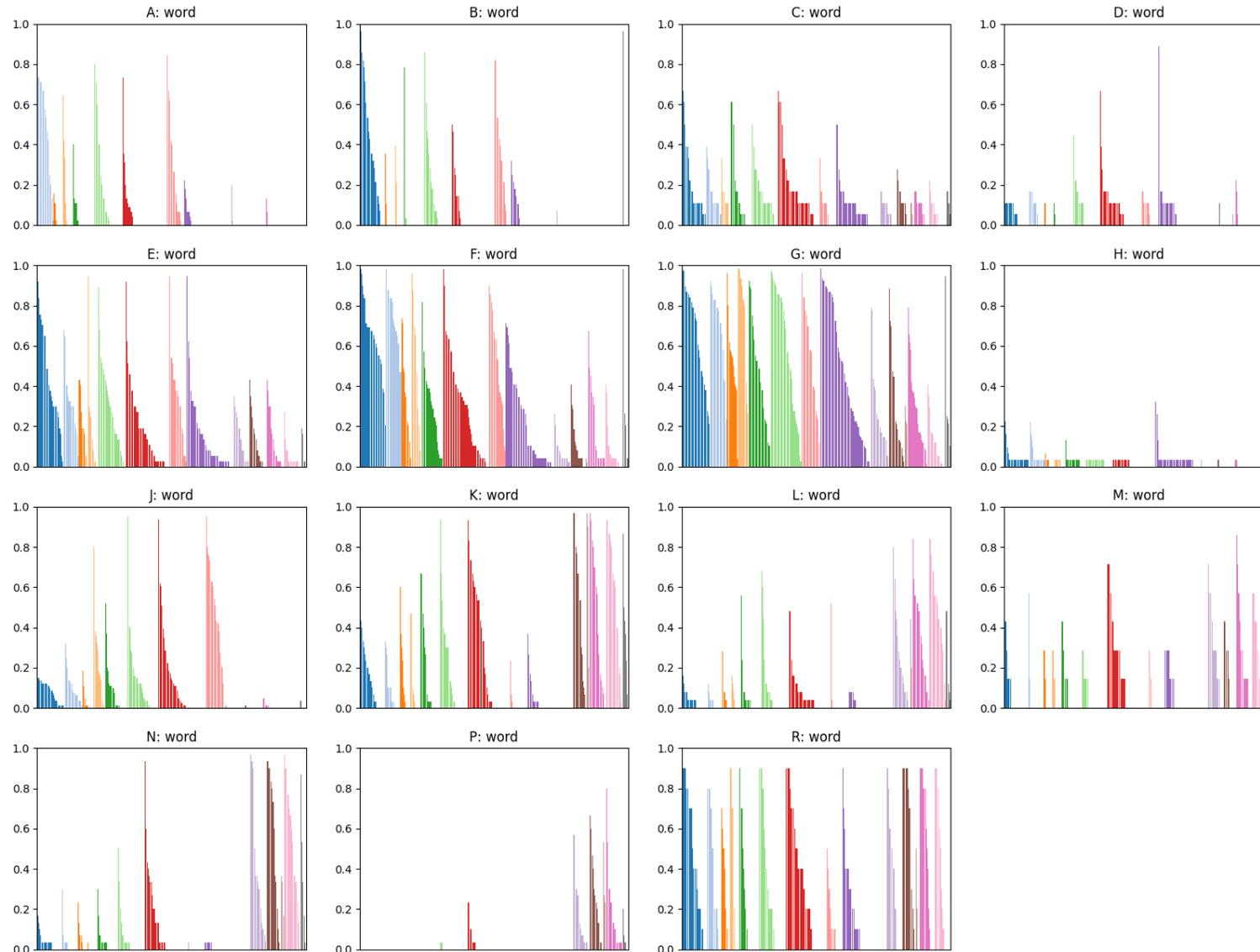
AmE06 categories + 1 outlier category text (POS)

Category	Register	Outlier File	Ranking	Total Files	Diff
A	Press: Reportage	K12	45	45	0
B	Press: Editorial	L9	25	28	3
C	Press: Reviews	P13	18	18	0
D	Religion	C8	10	18	8
E	Skills, Trades and Hobbies	N7	32	37	5
F	Popular Lore	A3	5	49	44
G	Belles Lettres, Biographies, Essays	M6	49	76	27
H	Miscellaneous: Government documents, industrial reports etc	L13	31	31	0
J	Academic prose in various disciplines	R8	80	81	1
K	General Fiction	E15	30	30	0
L	Mystery and Detective Fiction	C6	18	25	7
M	Science Fiction	N8	7	7	0
N	Adventure and Western	A7	30	30	0
P	Romance and Love story	A5	30	30	0
R	Humour	L2	3	10	7

Log Likelihood (LL2); $p < 0.05$

Experiment 3: Ranking of 'outlier' texts

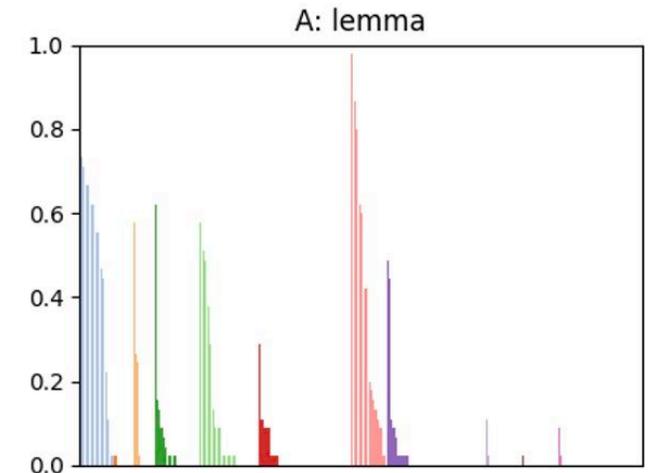
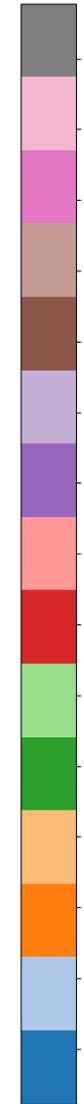
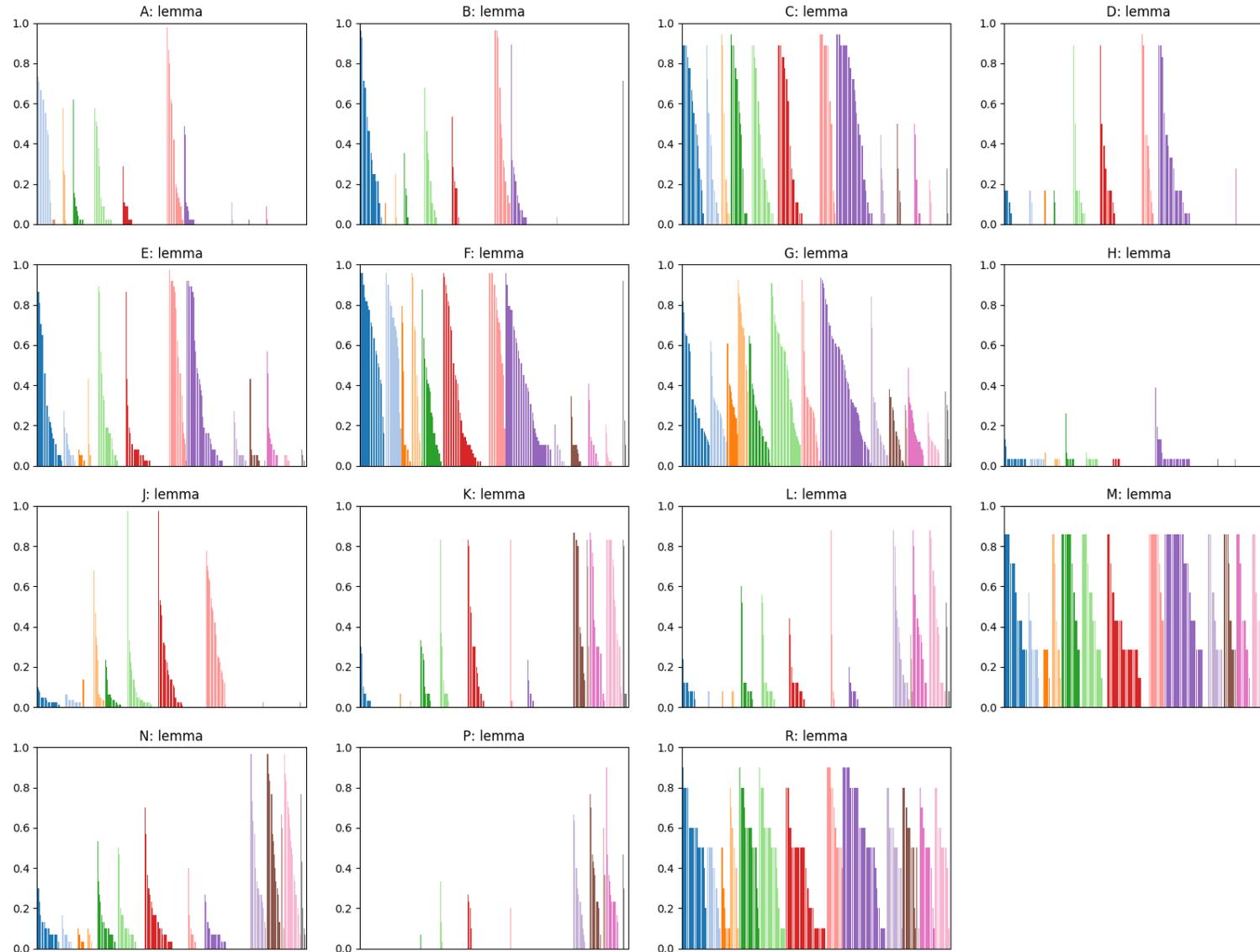
AmE06 categories + 1 outlier category text (LEX)



Log Likelihood (LL2); $p < 0.0001$

Experiment 3: Ranking of 'outlier' texts

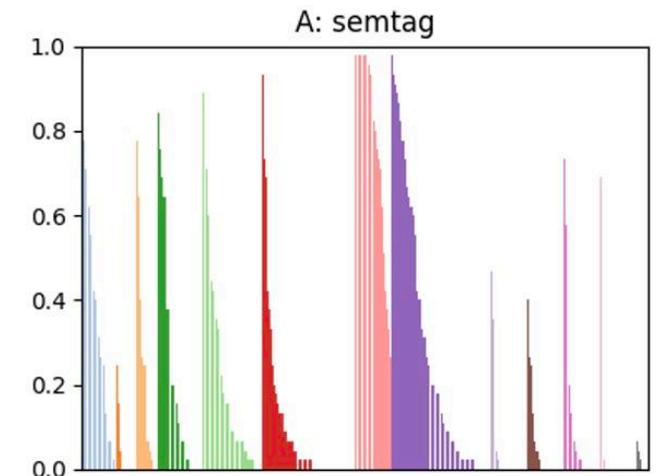
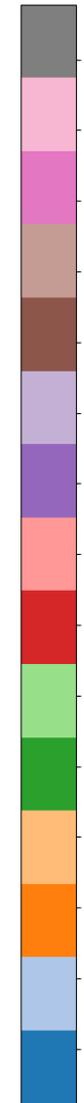
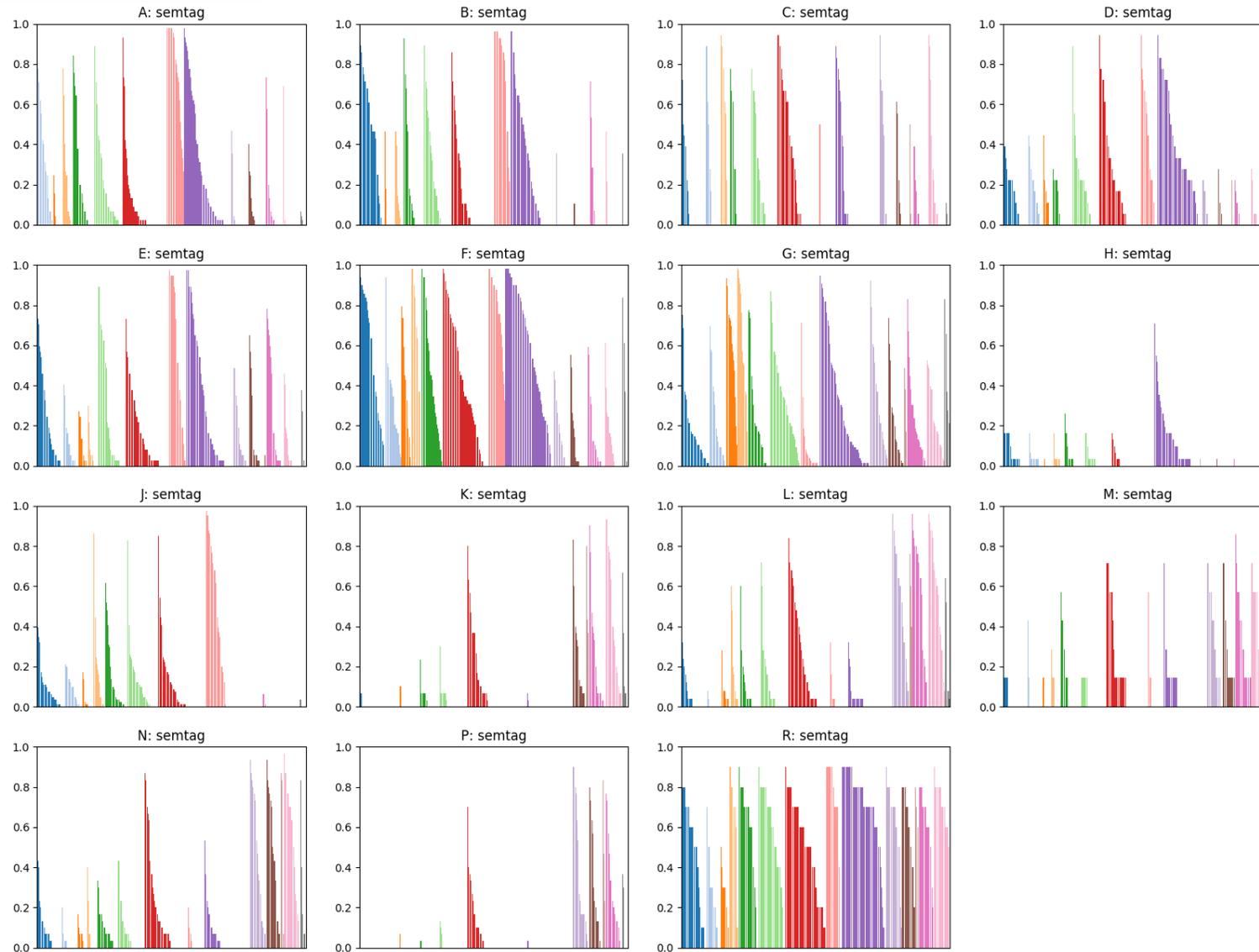
AmE06 categories + 1 outlier category text (LEM)



Log Likelihood (LL2); $p < 0.0001$

Experiment 3: Ranking of 'outlier' texts

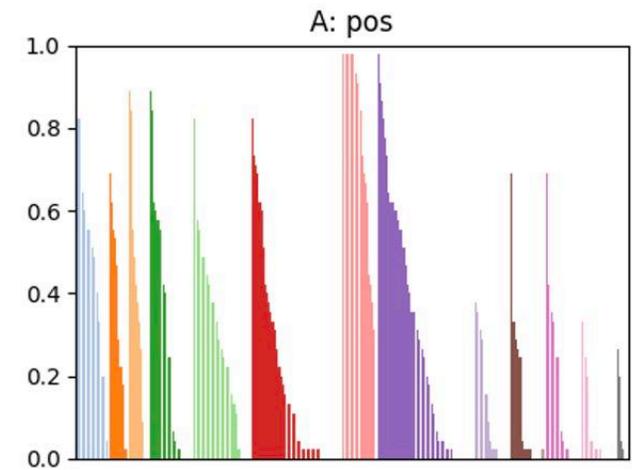
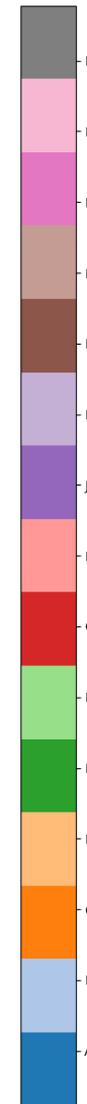
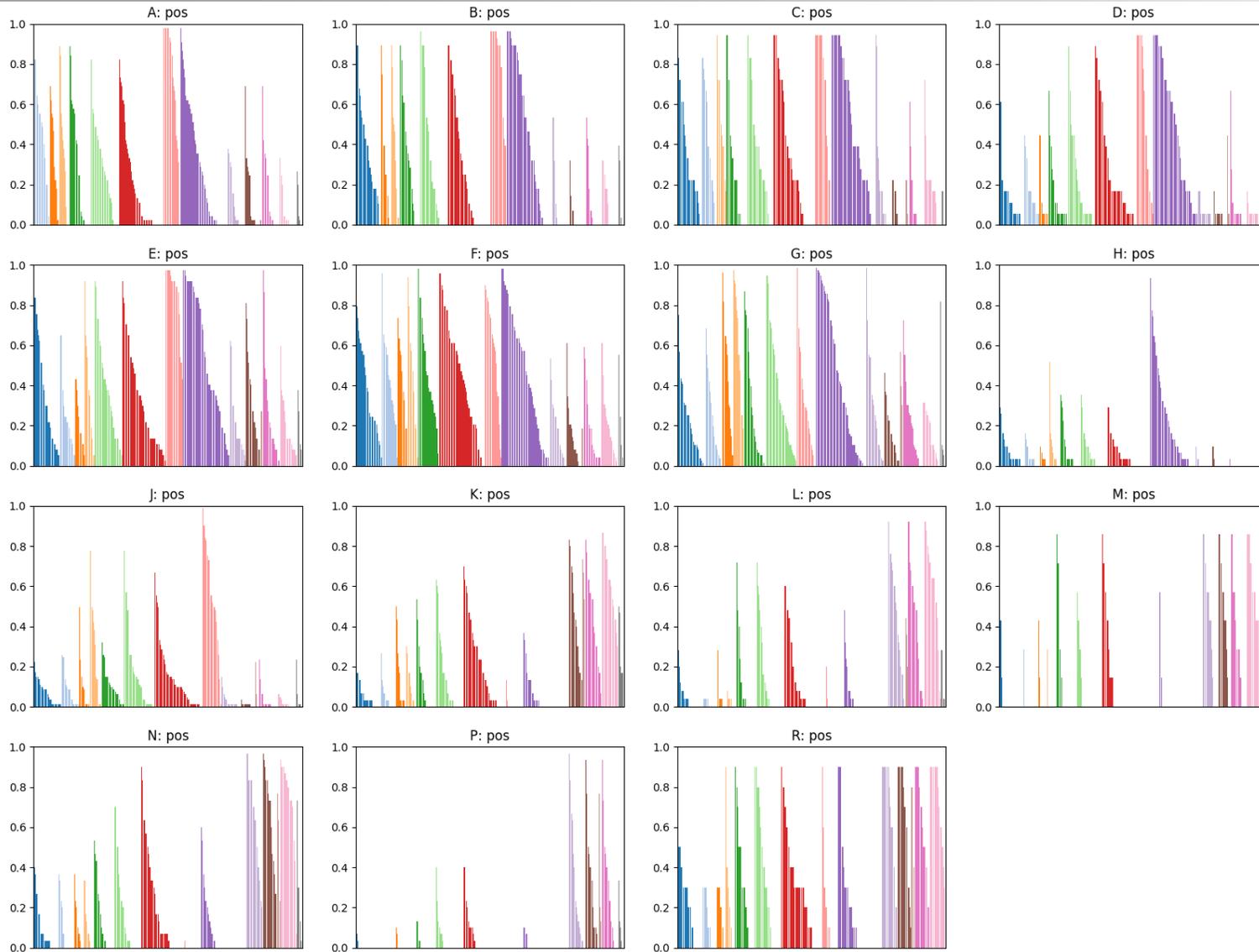
AmE06 categories + 1 outlier category text (SEM)



Log Likelihood (LL2); $p < 0.05$

Experiment 3: Ranking of 'outlier' texts

AmE06 categories + 1 outlier category text (POS)



Log Likelihood (LL2); $p < 0.05$

RQ2

How can ProtAnt be improved to allow more nuanced rankings?

How can *ProtAnt* be improved?

Ranking of 'outlier' texts (SEM)

Category	Register	Outlier File	Ranking	Total Files	Diff
A	Press: Reportage	K12	45	45	0
B	Press: Editorial	L9	28	28	0
C	Press: Reviews	P13	18	18	0
D	Religion	C8	14	18	4
E	Skills, Trades and Hobbies	N7	36	37	1
F	Popular Lore	A3	46	49	3
G	Belles Lettres, Biographies, Essays	M6	63	76	13
H	Miscellaneous: Government documents, industrial reports etc	L13	31	31	0
J	Academic prose in various disciplines	R8	81	81	0
K	General Fiction	E15	30	30	0
L	Mystery and Detective Fiction	C6	18	25	7
M	Science Fiction	N8	6	7	1
N	Adventure and Western	A7	30	30	0
P	Romance and Love story	A5	30	30	0
R	Humour	L2	3	10	7

Log Likelihood (LL2); $p < 0.05$

How can *ProtAnt* be improved?

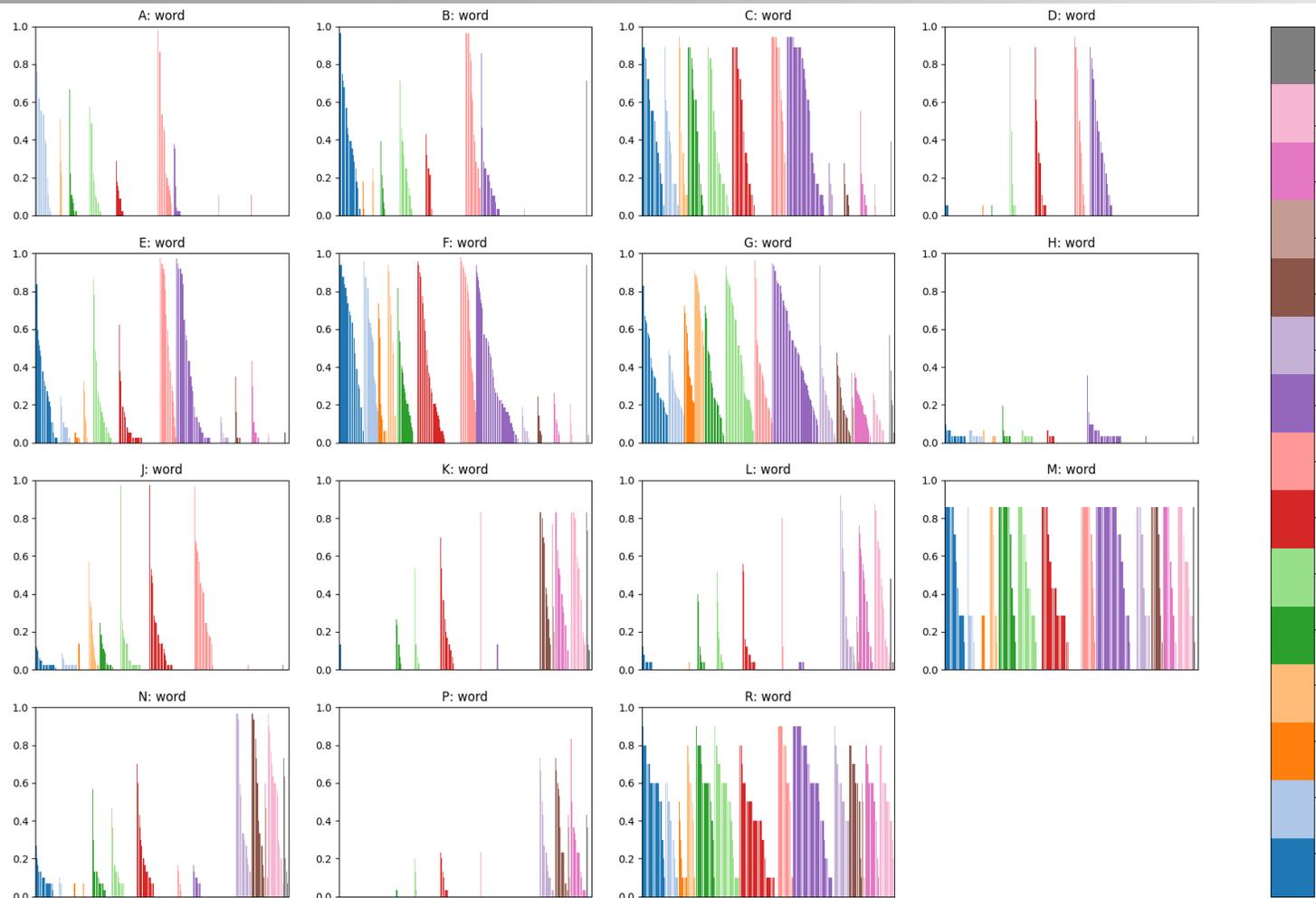
Ranking of 'outlier' texts (SEM) – SEM tag selection

Category	Register	Outlier File	Ranking	Total Files	Diff
A	Press: Reportage	K12	35	45	10
B	Press: Editorial	L9	28	28	0
C	Press: Reviews	P13	11	18	7
D	Religion	C8	18	18	0
E	Skills, Trades and Hobbies	N7	37	37	0
F	Popular Lore	A3	30	49	19
G	Belles Lettres, Biographies, Essays	M6	47	76	29
H	Miscellaneous: Government documents, industrial reports etc	L13	25	31	6
J	Academic prose in various disciplines	R8	51	81	30
K	General Fiction	E15	30	30	0
L	Mystery and Detective Fiction	C6	23	25	2
M	Science Fiction	N8	7	7	0
N	Adventure and Western	A7	27	30	3
P	Romance and Love story	A5	29	30	1
R	Humour	L2	7	10	3

Log Likelihood (LL2); $p < 0.05$; Highest SEM Key item only

How can *ProtAnt* be improved?

Ranking of 'outlier' texts – Using keyword text dispersion (LEX)

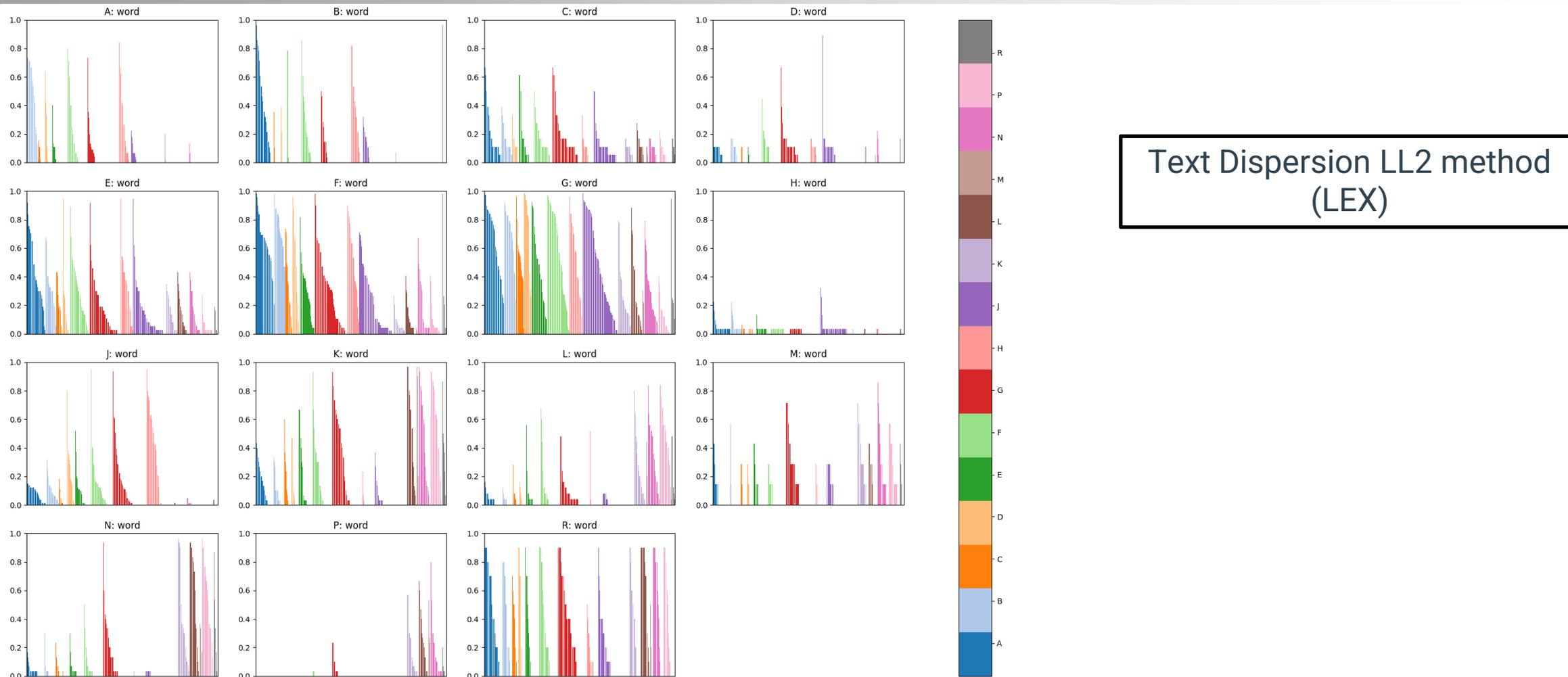


Standard LL2 method
(LEX)

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*. DOI:[10.3366/COR.2019.0162](https://doi.org/10.3366/COR.2019.0162)

How can *ProtAnt* be improved?

Ranking of 'outlier' texts – Using keyword text dispersion (LEX)



Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*. DOI: [10.3366/COR.2019.0162](https://doi.org/10.3366/COR.2019.0162)

RQ3

What are the implications of corpus prototypicality for corpus design and methods?

Implications for corpus design and methods

- A general assumption in corpus design is that text categories (topics, registers, genres, domains, etc.) are real
 - e.g., Brown/LOB family categories
 - A. PRESS: REPORTAGE (44 texts)
 - B. PRESS: EDITORIAL (27 texts)
 - C. PRESS: REVIEWS (17 texts)
 - D. RELIGION (17 texts)
 - E. SKILL AND HOBBIES (36 texts)
 - F. POPULAR LORE (48 texts)
 - G. BELLES-LETTRES (75 texts)
 - H. MISCELLANEOUS: GOVERNMENT & HOUSE ORGANS (30 texts)
 - J. LEARNED (80 texts)
 - K: FICTION: GENERAL (29 texts)
 - L: FICTION: MYSTERY (24 texts)
 - M: FICTION: SCIENCE (6 texts)
 - N: FICTION: ADVENTURE (29 texts)
 - P.FICTION: ROMANCE (29 texts)
 - R. HUMOR (9 texts)

Implications for corpus design and methods

- A general assumption in corpus design is that text categories (topics, registers, genres, domains, etc.) are real
 - e.g., Brown/LOB family categories
- How real are these categories?
 - What can we say about the characteristic features (LEXICAL, LEMMA, SEM, POS) of corpus text categories?
 - How much do the **individual texts** in corpus text categories match the category descriptions (at the LEXICAL, LEMMA, SEM, POS layers)?
- Our results suggest caution...
 - **"Remember the text!"** (Anthony, 2022)

Conclusions

- Replicating corpus studies is not easy and has many challenges...
 - locating the original texts
 - matching the experiment conditions
 - interpreting the results
- *ProtAnt* experiments here model prototypicality in the form of typicality/representativeness of LEX/LEM/SEM/POS forms
 - LEX/LEM rankings were generally 'stable' for different parameter settings
 - SEM/POS rankings required careful parameter selection
- *ProtAnt* can be improved in many ways...
 - allow for (easy) processing of texts in different annotation layers
 - allow more choices for keyword statistics and ranking measures (e.g., text dispersion)
 - allow more options for (batch) visualizations of the results
 - [offer ways to evaluate prototypicality of texts without a reference corpus]