Variation in morphological productivity in the history of English

The case of *-er*

Tanja Säily, Jukka Suomela & Eetu Mäkelä

UNIVERSITY OF HELSINKI

Aalto University

Nominal suffixes -*er* and -*or*

- Typically derive agentive or instrumental nouns from verbs: *driver*, *governor*, *filler*
  - Also: person ‘concerned with N’ / ‘living in N’
  - Here: division into **animate vs. inanimate**
- -*or*: Latinate variant of -*er*
  - Pronounced identically (*adviser*/*advisor*)
    → treated as a single suffix here
    (Plag 2003: 89; Bauer 2001: 199–203)

Productivity of -*er* and -*or*

- **Sociolinguistic variation**?
  - "Default" suffix (Bauer et al. 2013: 232) → no?
  - Säily (2011): women use -*ity* less productively than men; lower-class women: also -*ness*
    - Gendered styles?
- Productivity ≈ type frequency
  - Number of different words containing the suffix
  - Baayen (2009), Säily & Suomela (2009)

Material

- BNC = **British National Corpus**, early 1990s
  - Demographically sampled spoken component, both gender & social class known: 358 speakers, 2.6 Mw
- CEEC = **Corpora of Early English Correspondence**, C18 section (1680–1800)
  - Speech-like genre, social metadata
  - 315 writers, 2.2 Mw

Methods

- Types cross-checked with **MorphoQuantics** (Laws & Ryder 2014a, b)

- CEEC: **FiCa** interface for classifying data (developed by Eetu Mäkelä)

- Analysis of productivity: ***types*2** (Suomela 2014, 2015)

MorphoQuantics

-er*[3] *(Forms Ns from Adjs, Vs & Ns) ; Language of Origin = ***Germanic / Latin*** ; Etymon = ***-ârjoz / -ārius*** ; Meaning = ***an entity that performs the action specified by the verb*** ; Type = ***Suffix***

Export

Headword	PoS	Confirmed PoS	Tokens	Summed	Comment	Types
REDUCERS	NN2	n	2	2		1
REFRESHER	NN1	n	8	8		1
REINFORCER	NN1	n	2	3		1
REMINDER	NN1	n	54	62		1
RETAINER	NN1	n	5	5		1
REVIVER	NN1	n	1	1		1
REVOLVER	NN1	n	34	35		1
RINGER	NN1	n	3	3		1
ROASTER	NN1	n	1	2		1
ROCKER	NN1	n	4	4		1
ROLLER	NN1	n	88	112		1
ROLLER-COASTER	NN1	n	1	1		1
RUBBER	NN1	n	26	93		1
RULER	NN1	n	26	30		1
RUNNER	NN1	n	4	17		1
SANDER	NN1	n	8	10		1
SAVER	NN1	n	12	22		1
SCANNER	NN1	n	18	24		1
SCOOTER	NN1	n	29	33		1
SCORCHER	NN1	n	1	1		1
SCOURERS	NN2	n	1	1		1
SCRAPER	NN1	n	6	10		1
SCRAPPER	NN1	n	8	8		1
SCREWDRIVER	NN1	n	36	43		1
SCRUBBER	NN1	n	2	4		1
SEALER	NN1	n	7	8		1
SERVER	NN1	n	129	140		1
SERVERS	NN2	n	0	0		0
SHAKER	NN1	n	2	3		1
SHARPENER	NN1	n	12	17		1
SHAVER	NN1	n	5	7		1

FiCa

roller (3)					**RLR (3)**	**RLR (3)**
roller	rolers	yes		2	RLR	RLR
roller	roller	yes		3	RLR	RLR
roller	rollers	yes		6	RLR	RLR
ruler	ruler	unclear	er1; 1 in ed	5	**RLR**	**RLR**
rem~ber	rem~ber	no		2	**RMBR**	RMBR
					RMLR (2)	
					RMLR (2)	**RMLR (2)**
rambler (2)					**RMLR (2)**	**RMLR (2)**
rambler	rambler	unclear	some NP0	7	RMLR	RMLR
rambler	ramblers	no	NP0 (perioc	1	RMLR	RMLR
remem[{b{]er	remem[{b{]e	no		1	**RMMB**	RMMBR
					RMMR (8)	
					RMMR (7)	**RMMR (7)**
rememb=r=s	rememb=r=s	no		1	**RMMR**	**RMMR**
remember (2)					**RMMR (2)**	**RMMR (2)**
remember	rememb=r=	no		8	RMMR	RMMR
remember	remember	no		690	RMMR	RMMR
remember's	remember's	no		2	**RMMR**	**RMMR**
remembers	remembers	no		22	**RMMR**	**RMMR**
remembr	remembr	no		2	**RMMR**	**RMMR**
remmember	remmember	no		2	**RMMR**	**RMMR**
remembrancer	remembranc	yes	er1	1	**RMMR**	**RMMRNSR**
					RMNT (9)	
					RMNT (9)	**RMNTR (9)**
remainder (9)					**RMNT (9)**	**RMNTR (9)**
remainder	rem[]ainder	no	MO	1	RMNT	RMNTR

of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone **rolers:** this he says is a late invention; but as it requires more workmanship than is easy to be had here I think cast iron rol

odel he has brought with him from England there is an apparatus for bruising the seed by making it pass between two iron **rolers** of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone

<Q A 1783 FN SBENTHAM> <X SAMUEL BENTHAM> <P III,209> [] [\469 FROM SAMUEL BENTHAM\]]] [^TO JEREMY BENTHAM^] [\ADDRESSED:\] Jere=y= Bentham Esq=r= / Lincoln's Inn / London Petersbourg Septr. 13th O.S. 1783. I am at length taken into the service of this country. The rank given me is that of (\Conseiller de la Cour\) , which is only equal to that of Lieutenant Colonel in the army. Considering that I had had no Military rank in any other country to found my pretentions on, ...^) to bruise the seed in oil mills, and what objections he sees to such a substitute. In many places I imagine the rollers of stone may be cheaper; but that would not be the case here. Capper is returned and his having seen you all has attatched me to him. It is probable we may be concerned together in the erection of an oil mill which gave rise to the above question. In a model he has brought with him from England there is an apparatus for bruising the seed by making it pass between two iron rolers of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone

types2

A tool for exploring word-frequency differences in corpora

Comparing word frequencies

- Type frequency = extent of use or realised productivity (Baayen 2009)
  - Cannot be normalised → difficult to compare subcorpora, e.g. different social groups
- *types*2: **permutation testing**
  - Compare single subcorpus with multiple randomly composed subcorpora of the same size
  - Random subcorpora sampled from the entire corpus → represent what is normal in it

Exploring word frequencies

- Typically: static tables, figures
  - Not conducive to rapid exploration
- **Interpretation** of results?
  - Need to go back to the concordances & metadata
- *types*2: online interface with interactive figures, **linked data**

Case 1: BNC

Demographically sampled spoken component, early 1990s

types2

Overview
Plot
Types
Samples
Help

Corpus
bnc-spoken-demo
bnc-spoken-demo-home

Dataset
er+or
er+or person
er+or thing

Group
all
age
age + gender
gender
social class
social class + gender
none

Collection
none

Statistics
types / running words
types / tokens

corpus	dataset	collection	axes	side	p-value	FDR
bnc-spoken-demo	er+or thing	Female	types / running words	below	0.000016	0.0063
bnc-spoken-demo	er+or	Female	types / running words	below	0.000022	0.0063
bnc-spoken-demo	er+or thing	Male	types / running words	above	0.00015	0.020
bnc-spoken-demo	er+or thing	Female	types / tokens	below	0.00036	0.037
bnc-spoken-demo	er+or	Male	types / running words	above	0.00079	0.064
bnc-spoken-demo-home	er+or	45- Male	types / running words	above	0.00081	0.064
bnc-spoken-demo-home	er+or	Female	types / running words	below	0.00083	0.064
bnc-spoken-demo	er+or person	C2+DE Female	types / running words	below	0.0013	0.064
bnc-spoken-demo-home	er+or thing	Female	types / running words	below	0.0020	0.090
bnc-spoken-demo	er+or	C2+DE Female	types / running words	below	0.0021	0.090
bnc-spoken-demo-home	er+or	Male	types / running words	above	0.0023	0.090
bnc-spoken-demo	er+or thing	Male	types / tokens	above	0.0026	0.090
bnc-spoken-demo	er+or person	Female	types / running words	below	0.0029	0.092
bnc-spoken-demo-home	er+or person	C2+DE Female	types / running words	below	0.0034	0.10
bnc-spoken-demo-home	er+or person	45- Male	types / running words	above	0.0039	0.10
bnc-spoken-demo	er+or thing	45- Male	types / tokens	above	0.0045	0.11
bnc-spoken-demo	er+or thing	C2+DE Male	types / running words	above	0.0046	0.11
bnc-spoken-demo-home	er+or person	Female	types / running words	below	0.0057	0.13
bnc-spoken-demo	er+or thing	45- Female	types / running words	below	0.0057	0.13
bnc-spoken-demo-home	er+or	C2+DE Male	types / running words	above	0.0059	0.13
bnc-spoken-demo-home	er+or thing	AB+C1 Female	types / running words	below	0.0064	0.13
bnc-spoken-demo	er+or thing	AB+C1 Female	types / running words	below	0.0065	0.13

The corpus **bnc-spoken-demo** contains 358 samples and 2,632,512 running words.

The dataset **er+or** contains 249 hapaxes, 6,431 types, and 692 tokens.

dataset	collection	axes	side	p-value	FDR
er+or thing	Female	types / running words	below	0.000016	0.0063
er+or	Female	types / running words	below	0.000022	0.0063
er+or thing	Male	types / running words	above	0.00015	0.020
er+or thing	Female	types / tokens	below	0.00036	0.037
er+or	Male	types / running words	above	0.00079	0.064
er+or	45- Male	types / running words	above	0.00081	0.064
er+or	Female	types / running words	below	0.00083	0.064
er+or person	C2+DE Female	types / running words	below	0.0013	0.064
er+or thing	Female	types / running words	below	0.0020	0.090
er+or	C2+DE Female	types / running words	below	0.0021	0.090
er+or	Male	types / running words	above	0.0023	0.090
er+or thing	Male	types / tokens	above	0.0026	0.090
er+or person	Female	types / running words	below	0.0029	0.092
er+or person	C2+DE Female	types / running words	below	0.0034	0.10
er+or person	45- Male	types / running words	above	0.0039	0.10
er+or thing	45- Male	types / tokens	above	0.0045	0.11
er+or thing	C2+DE Male	types / running words	above	0.0046	0.11
er+or person	Female	types / running words	below	0.0057	0.13

types2

- Overview
- **Plot**
- Types
- Samples
- Help

Corpus

- **bnc-spoken-demo**
- bnc-spoken-demo-home

Dataset

- er+or
- er+or person
- **er+or thing**

Group

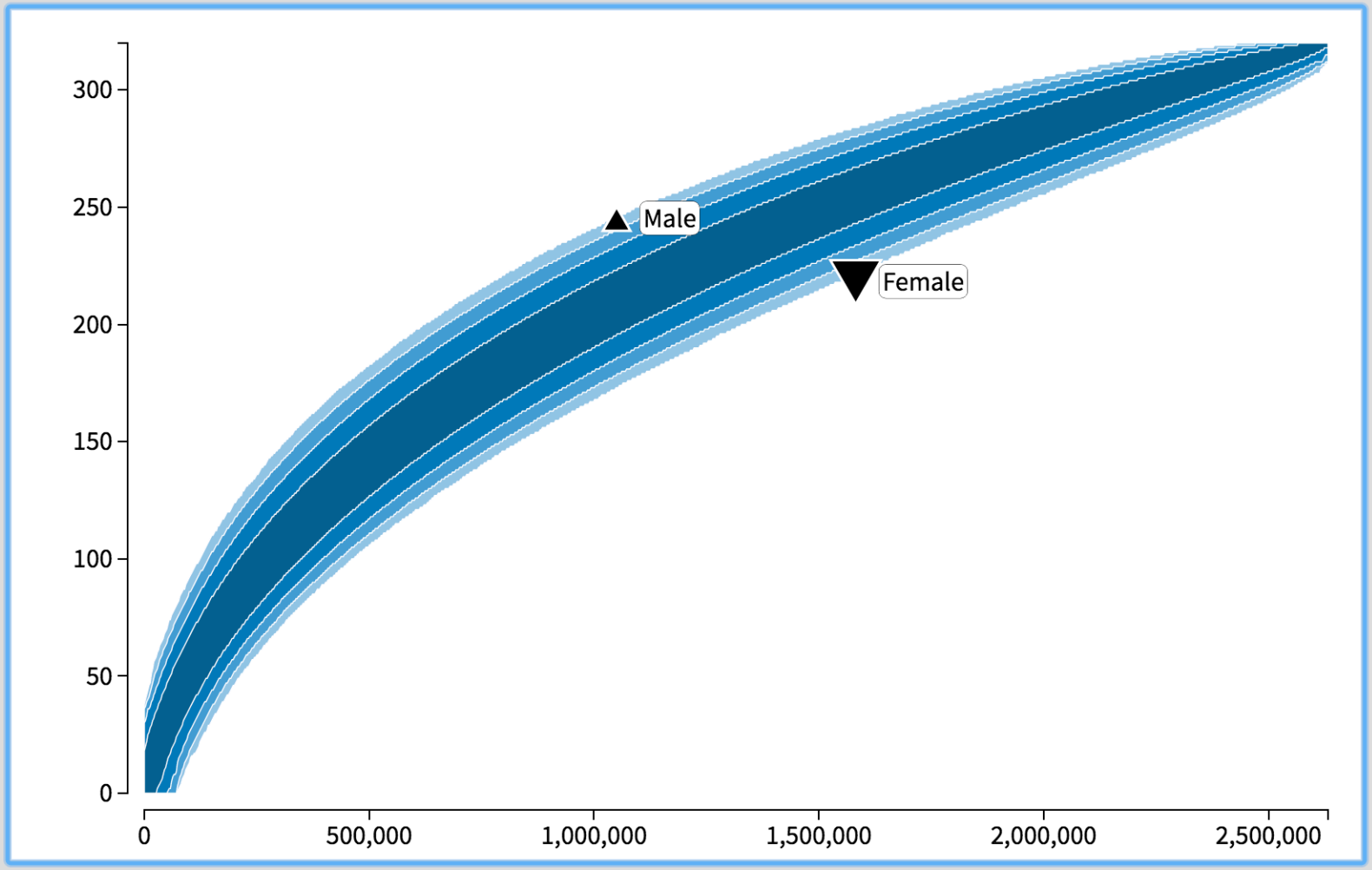
- all
- age
- age + gender
- **gender**
- social class
- social class + gender
- none

Collection

- none
- **Female**
- Male

Statistics

- **types / running words**
- types / tokens



The corpus **bnc-spoken-demo** contains 358 samples and 2,632,512 running words.

The dataset **er+or thing** contains 111 hapaxes, 3,318 types, and 320 tokens.

The collection **Female** contains 1,582,116 running words and 218 types.

Only **0.0016%** of random collections with 1,582,116 running words contain at most 218 types.

This finding is probably interesting: the false discovery rate is **0.0063**.

types2

Overview
Plot
Types
Samples
Help

Corpus

bnc-spoken-demo
bnc-spoken-demo-home

Dataset

er+or
er+or person
er+or thing

Group

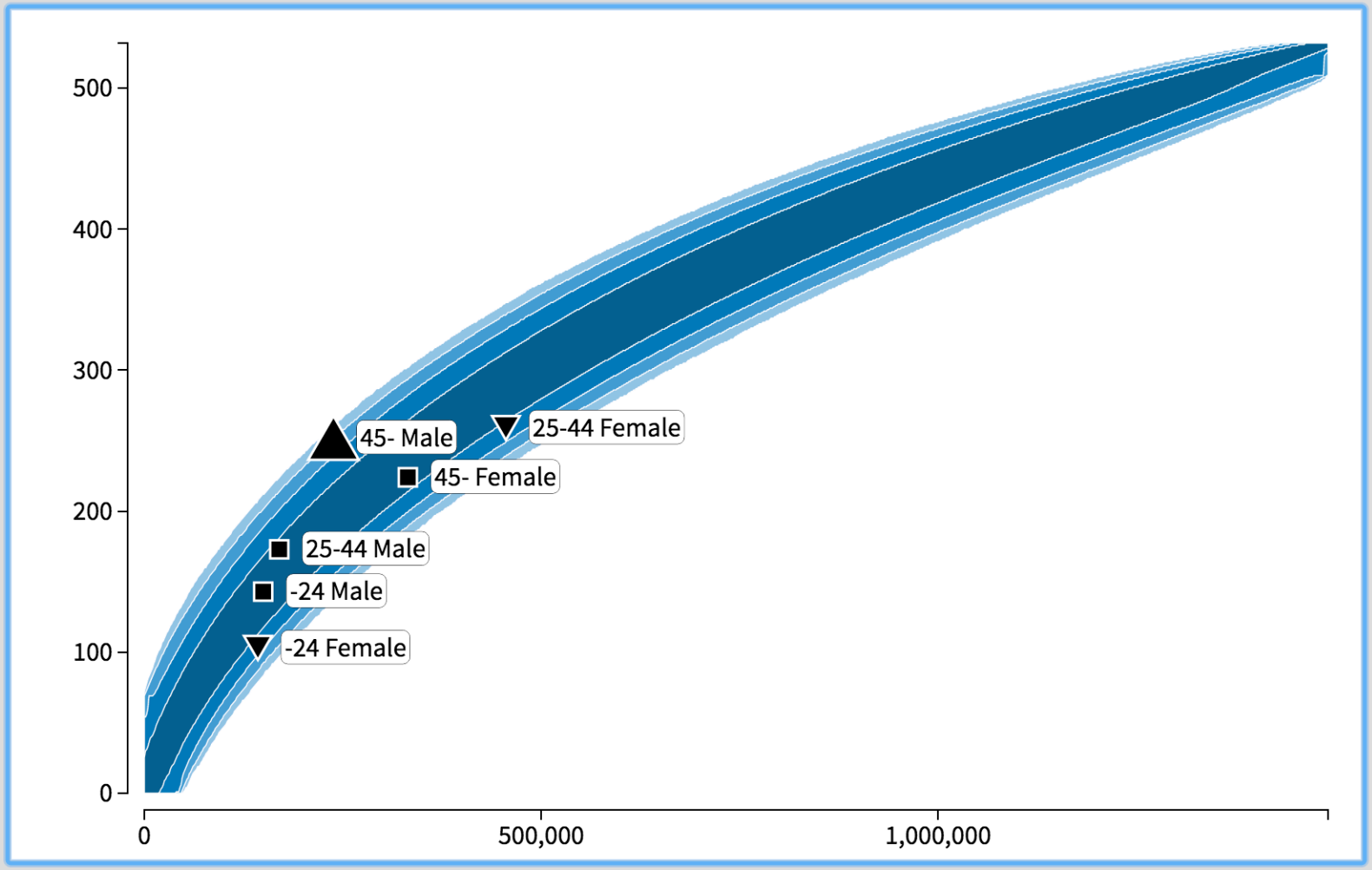
all
age
age + gender
gender
social class
social class + gender
none

Collection

none
-24 Female
-24 Male
25-44 Female
25-44 Male
45- Female
45- Male

Statistics

types / running words
types / tokens



The corpus **bnc-spoken-demo-home** contains 284 samples and 1,491,358 running words.

The dataset **er+or** contains 215 hapaxes, 3,648 types, and 532 tokens.

The collection **45- Male** contains 238,311 running words and 252 types.

Only **0.081%** of random collections with 238,311 running words contain at least 252 types.

This finding is probably interesting: the false discovery rate is **0.064**.

types2

- Overview
- **Plot**
- Types
- Samples
- Help

Corpus

- bnc-spoken-demo
- **bnc-spoken-demo-home**

Dataset

- er+or
- **er+or person**
- er+or thing

Group

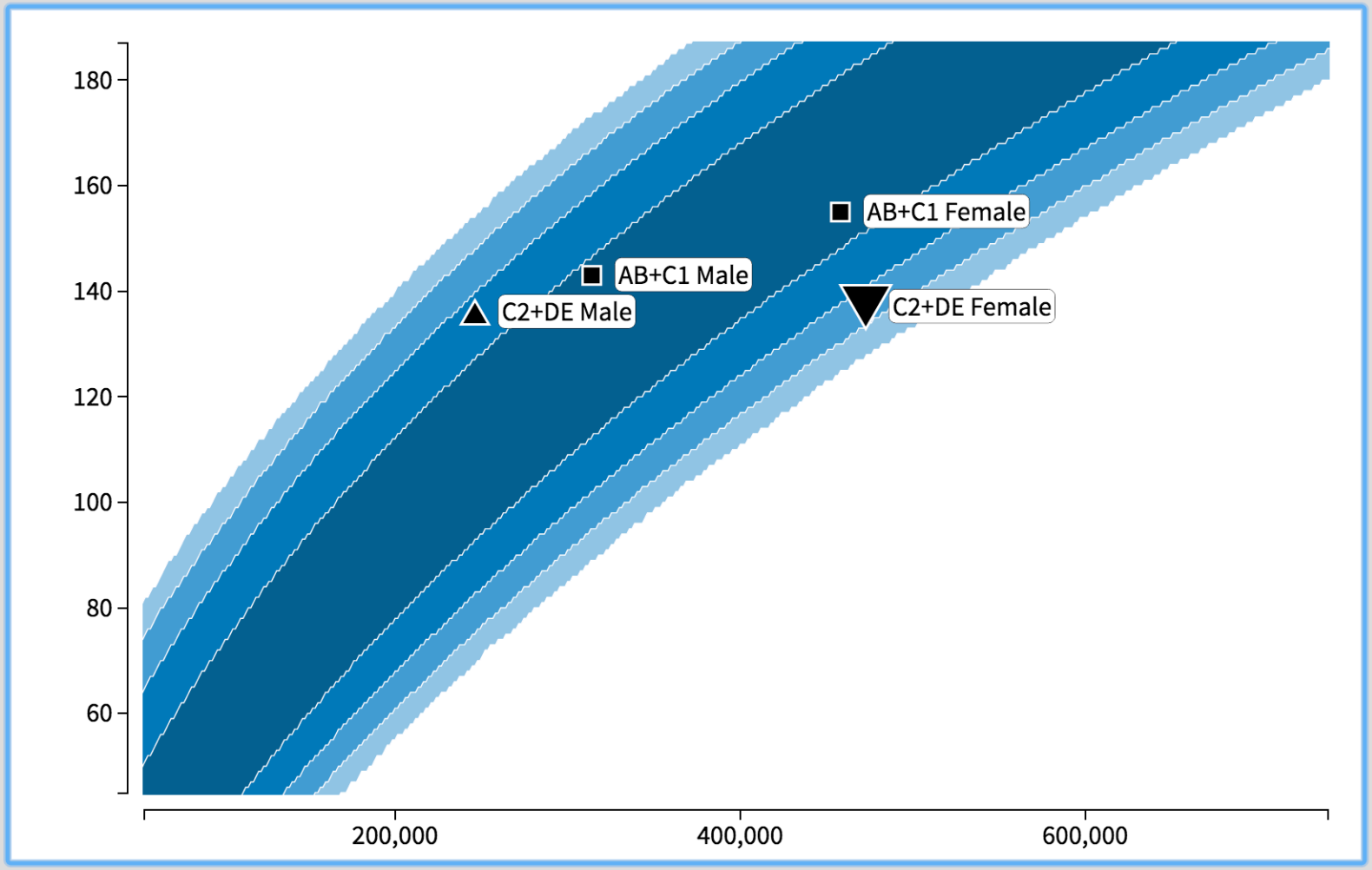
- all
- age
- age + gender
- gender
- social class
- **social class + gender**
- none

Collection

- none
- AB+C1 Female
- AB+C1 Male
- **C2+DE Female**
- C2+DE Male

Statistics

- **types / running words**
- types / tokens



The corpus **bnc-spoken-demo-home** contains 284 samples and 1,491,358 running words.

The dataset **er+or person** contains 130 hapaxes, 1,744 types, and 309 tokens.

The collection **C2+DE Female** contains 472,852 running words and 137 types.

Only **0.34%** of random collections with 472,852 running words contain at most 137 types.

This finding is probably not interesting: the false discovery rate is larger than 0.1.

BNC: Summary of results

- Men use -*er* more productively than women
- Especially older men, even at home
- Working-class women underuse animate -*er*

Women's use of inanimate *-er*

type	tokens	in collection	fraction	score	samples
dryer	31	24	0.774		20
duster	12	11	0.917		9
layer	13	10	0.769		9
breather	3	3	1.000		3
caster	4	4	1.000		3
downer	4	4	1.000		3
freshener	4	4	1.000		3
hopper	3	3	1.000		3
inhaler	5	5	1.000		3
refresher	3	3	1.000		3
whopper	3	3	1.000		3
bloomer	2	2	1.000		2

Men's use of inanimate *-er*

type	tokens	in collection	fraction	score	samples
bomber	12	12	1.000		7
breaker	4	4	1.000		3
generator	3	3	1.000		3
trimmer	3	3	1.000		3
blaster	2	2	1.000		2
booster	2	2	1.000		2
bowler	3	3	1.000		2
carver	3	3	1.000		2
cursor	2	2	1.000		2
decoder	8	8	1.000		2
edger	2	2	1.000		2
elevator	4	4	1.000		2

Who are the male users?

sample	description	running words	tokens	/1000
PS03S	C2 m 60- retired (precision engineer)	36,388	43	1.2
PS09E	C2 m 45-59 tv engineer	9,541	22	2.3
PS0H9	C2 m 35-44 telecommunication engineer	17,768	107	6.0
PS0JX	C1 m 25-34 technician	19,701	24	1.2
PS1GE	C2 m 25-34 aircraft engineer	23,488	21	0.89
PS0PN	C1 m 45-59 retired	36,385	44	1.2

Older men at home

sample	description	running words	tokens	/1000
PS03S	C2 m 60- retired (precision engineer)	36,388	43	1.2
PS0W2	AB m 60- retired (headteacher)	30,666	26	0.85
PS1BT	AB m 45-59 chartered engineer	11,813	24	2.0
PS05X	AB m 45-59 export merchant	16,499	18	1.1
PS065	DE m 60- retired	13,569	49	3.6
PS0PN	C1 m 45-59 retired	6,030	9	1.5
PS007	AB m 60- retired	1,187	4	3.4

Men's use of animate *-er*

type	tokens	in collection	fraction	score	samples
blighter	4	4	1.000		4
carpenter	4	4	1.000		4
hunter	4	4	1.000		4
listener	4	4	1.000		4
loser	11	11	1.000		4
banker	5	5	1.000		3
golfer	3	3	1.000		3
interviewer	5	5	1.000		3
tailor	6	6	1.000		3
tanner	3	3	1.000		3
treasurer	5	5	1.000		3
warrior	6	6	1.000		3

buster	9	8	0.889	6
defender	3	3	1.000	2
drinker	2	2	1.000	2
gambler	2	2	1.000	2
goner	2	2	1.000	2
investigator	3	3	1.000	2
kidnapper	4	4	1.000	2
lender	2	2	1.000	2
pacer	2	2	1.000	2

sample	description	before
PS09T	DE m 60- retired	poor old
PS0HM	C1 m 35-44 draughtsman	Cheeky
PS14C	AB m 45-59 charge nurse	Well these poor
PS4YX	AB m -14 student	had a history with his toe poor

BNC: Interpretation of results

- Men use -*er* more productively than women
  - Focus on tools & occupations, playful name-calling: masculine identity-building?
- Especially older men, even at home
  - Keune et al. (2006, 2012): highly educated older men are the most productive users of Dutch affixes
  - Štekauer et al. (2005): highly educated older speakers prefer more explicit naming strategies in English
- Working-class women underuse animate -*er*
  - Prefer other strategies of referring to people?
  - Involved style? → fewer nouns (cf. Säily 2011)

Case 2: CEEC

Corpora of Early English Correspondence, 1680–1800 (pilot results, handle with care!)

corpus	dataset	collection	axes	side	p-value	FDR
ceec-1680-1800	er+or person	Rank R	types / running words	below	0.000010	0.0029
ceec-1680-1800	er+or	Rank R	types / running words	below	0.000061	0.0088
ceec-1680-1800	er+or person	Female FN	types / running words	below	0.00018	0.017
ceec-1680-1800	er+or	Female FN	types / running words	below	0.00023	0.017
ceec-1680-1800	er+or	1720-1759	types / running words	below	0.00026	0.017
ceec-1680-1800	er+or	Male TC	types / running words	above	0.00027	0.017
ceec-1680-1800	er+or person	1680-1719	types / tokens	below	0.00034	0.017
ceec-1680-1800	er+or	1680-1719	types / tokens	below	0.00039	0.017
ceec-1680-1800	er+or	Female	types / running words	below	0.0011	0.034
ceec-1680-1800	er+or person	1720-1759	types / running words	below	0.0013	0.037
ceec-1680-1800	er+or person	Male TC	types / running words	above	0.0014	0.037
ceec-1680-1800	er+or	Rank C	types / tokens	below	0.0018	0.043
ceec-1680-1800	er+or person	Female	types / running words	below	0.0018	0.043
ceec-1680-1800	er+or	1720-1759	types / tokens	below	0.0026	0.054
ceec-1680-1800	er+or person	FN	types / running words	below	0.0033	0.064
ceec-1680-1800	er+or	Rank N	types / running words	below	0.0038	0.068
ceec-1680-1800	er+or person	Rank C	types / tokens	below	0.0048	0.081
ceec-1680-1800	er+or person	1720-1759	types / tokens	below	0.0057	0.091
ceec-1680-1800	er+or	FN	types / running words	below	0.0061	0.092
ceec-1680-1800	er+or person	Rank N	types / running words	below	0.0085	0.12
ceec-1680-1800	er+or thing	1720-1759	types / running words	below	0.0096	0.13
ceec-1680-1800	er+or	T	types / tokens	below	0.012	0.15
ceec-1680-1800	er+or	Male TC	types / tokens	above	0.012	0.16
ceec-1680-1800	er+or	Male T	types / tokens	below	0.015	0.19

The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or** contains 249 hapaxes, 607 types, and 5,318 tokens.

	dataset	collection	axes	side	p-value	FDR
-1800	er+or person	Rank R	types / running words	below	0.000010	0.0029
-1800	er+or	Rank R	types / running words	below	0.000061	0.0088
-1800	er+or person	Female FN	types / running words	below	0.00018	0.017
-1800	er+or	Female FN	types / running words	below	0.00023	0.017
-1800	er+or	1720-1759	types / running words	below	0.00026	0.017
-1800	er+or	Male TC	types / running words	above	0.00027	0.017
-1800	er+or person	1680-1719	types / tokens	below	0.00034	0.017
-1800	er+or	1680-1719	types / tokens	below	0.00039	0.017
-1800	er+or	Female	types / running words	below	0.0011	0.034
-1800	er+or person	1720-1759	types / running words	below	0.0013	0.037
-1800	er+or person	Male TC	types / running words	above	0.0014	0.037
-1800	er+or	Rank C	types / tokens	below	0.0018	0.043
-1800	er+or person	Female	types / running words	below	0.0018	0.043
-1800	er+or	1720-1759	types / tokens	below	0.0026	0.054
-1800	er+or person	FN	types / running words	below	0.0033	0.064
-1800	er+or	Rank N	types / running words	below	0.0038	0.068
-1800	er+or person	Rank C	types / tokens	below	0.0048	0.081
-1800	er+or person	1720-1759	types / tokens	below	0.0057	0.091
-1800	er+or	FN	types / running words	below	0.0061	0.092

types2

- Overview
- **Plot**
- Types
- Samples
- Help

Corpus

- **ceec-1680-1800**

Dataset

- er+or
- **er+or person**
- er+or thing

Group

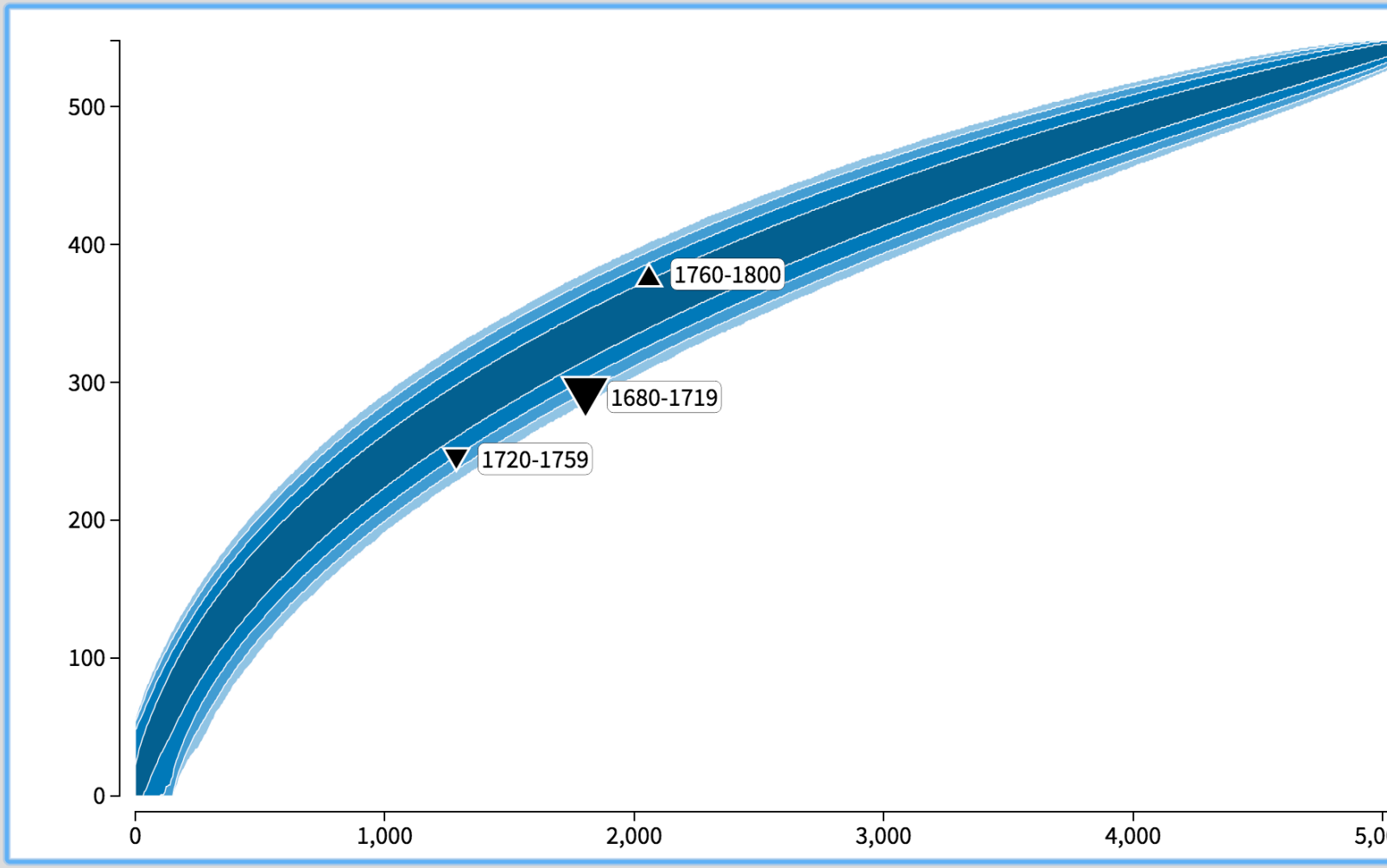
- all
- gender
- gender + relcode
- **period**
- rank
- relcode
- none

Collection

- none
- **1680-1719**
- 1720-1759
- 1760-1800

Statistics

- types / running words
- **types / tokens**



The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or person** contains 220 hapaxes, 548 types, and 5,151 tokens.

The collection **1680-1719** contains 1,805 tokens and 289 types.

Only **0.034%** of random collections with 1,805 tokens contain at most 289 types.

This finding is probably interesting: the false discovery rate is **0.017**.

types2

Overview
Plot
Types
Samples
Help

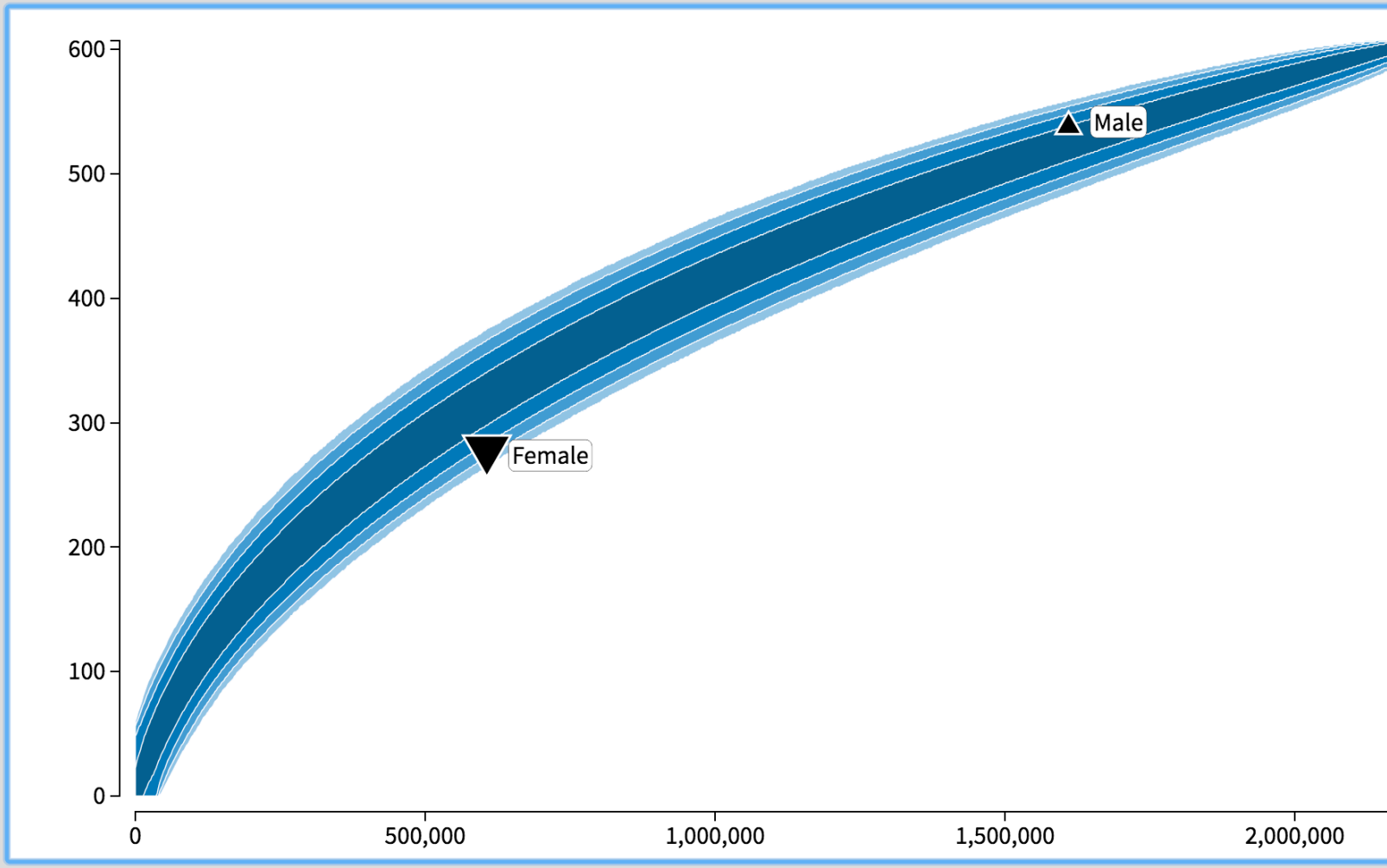
Corpus
ceec-1680-1800

Dataset
er+or
er+or person
er+or thing

Group
all
gender
gender + relcode
period
rank
relcode
none

Collection
none
Female
Male

Statistics
types / running words
types / tokens



The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or** contains 249 hapaxes, 607 types, and 5,318 tokens.

The collection **Female** contains 606,366 running words and 273 types.

Only **0.11%** of random collections with 606,366 running words contain at most 273 types.

This finding is probably interesting: the false discovery rate is **0.034**.

types2

- Overview
- **Plot**
- Types
- Samples
- Help

Corpus

- **ceec-1680-1800**

Dataset

- **er+or**
- er+or person
- er+or thing

Group

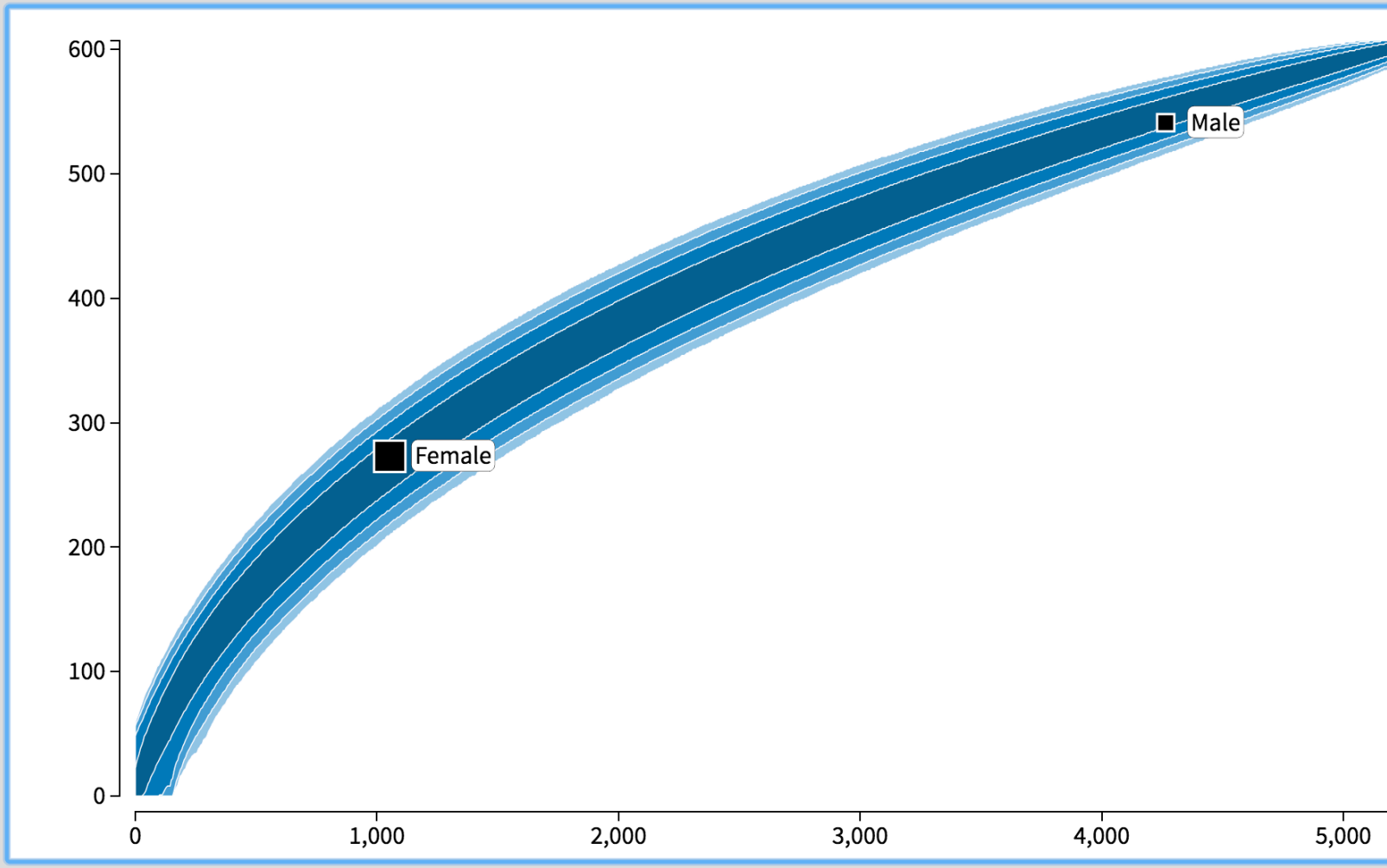
- all
- **gender**
- gender + relcode
- period
- rank
- relcode
- none

Collection

- none
- **Female**
- Male

Statistics

- types / running words
- **types / tokens**



The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or** contains 249 hapaxes, 607 types, and 5,318 tokens.

The collection **Female** contains 1,053 tokens and 273 types.

Approximately **39%** of random collections with 1,053 tokens contain at least 273 types.

This seems to be a fairly typical collection.

types2

- Overview
- **Plot**
- Types
- Samples
- Help

Corpus

- **ceec-1680-1800**

Dataset

- **er+or**
- er+or person
- er+or thing

Group

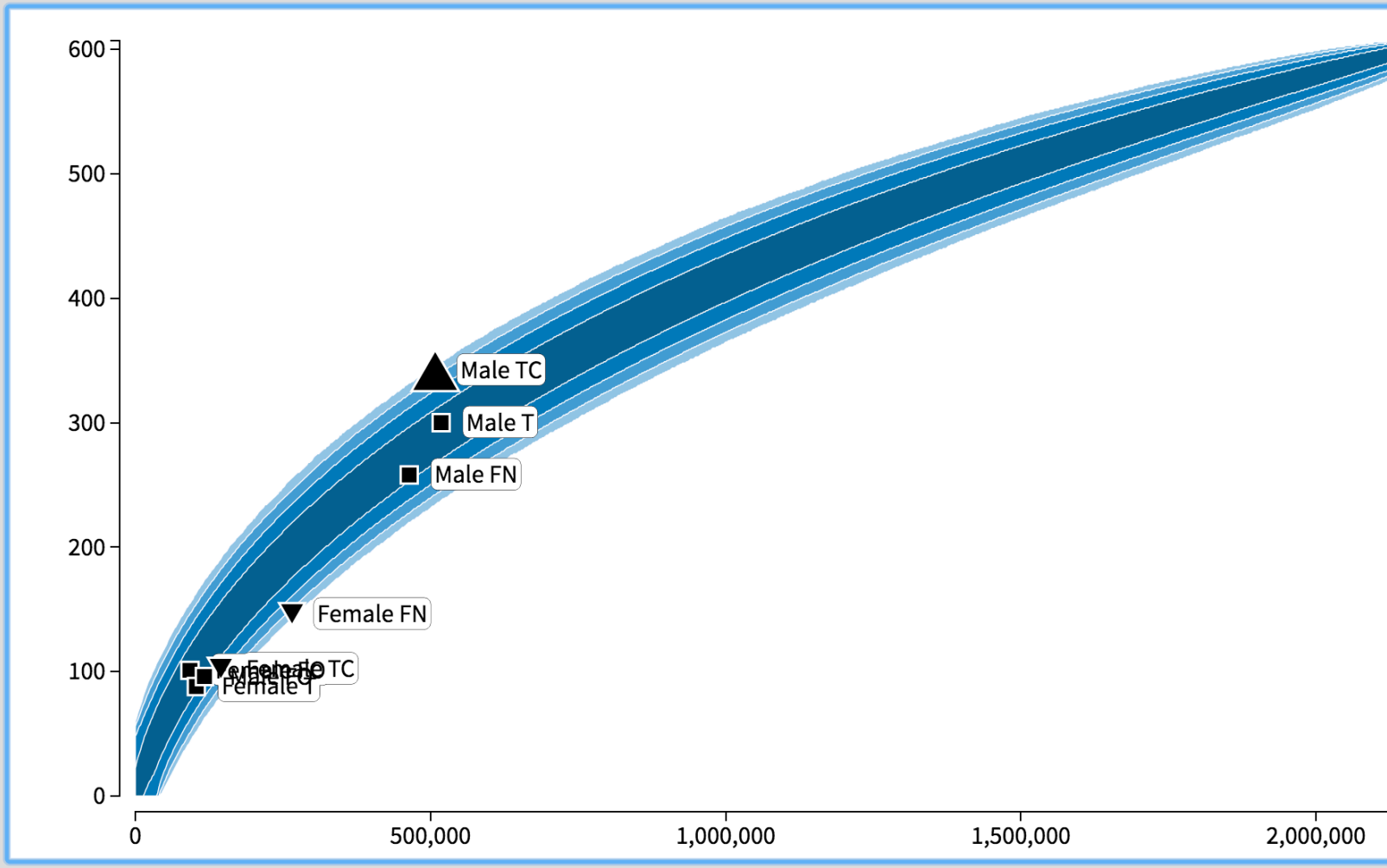
- all
- gender
- **gender + relcode**
- period
- rank
- relcode
- none

Collection

- none
- Female FN
- Female FO
- Female T
- Female TC
- Male FN
- Male FO
- Male T
- **Male TC**

Statistics

- **types / running words**
- types / tokens



The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or** contains 249 hapaxes, 607 types, and 5,318 tokens.

The collection **Male TC** contains 507,958 running words and 342 types.

Only **0.027%** of random collections with 507,958 running words contain at least 342 types.

This finding is probably interesting: the false discovery rate is **0.017**.

types2

- Overview
- **Plot**
- Types
- Samples
- Help

Corpus

- **ceec-1680-1800**

Dataset

- **er+or**
- er+or person
- er+or thing

Group

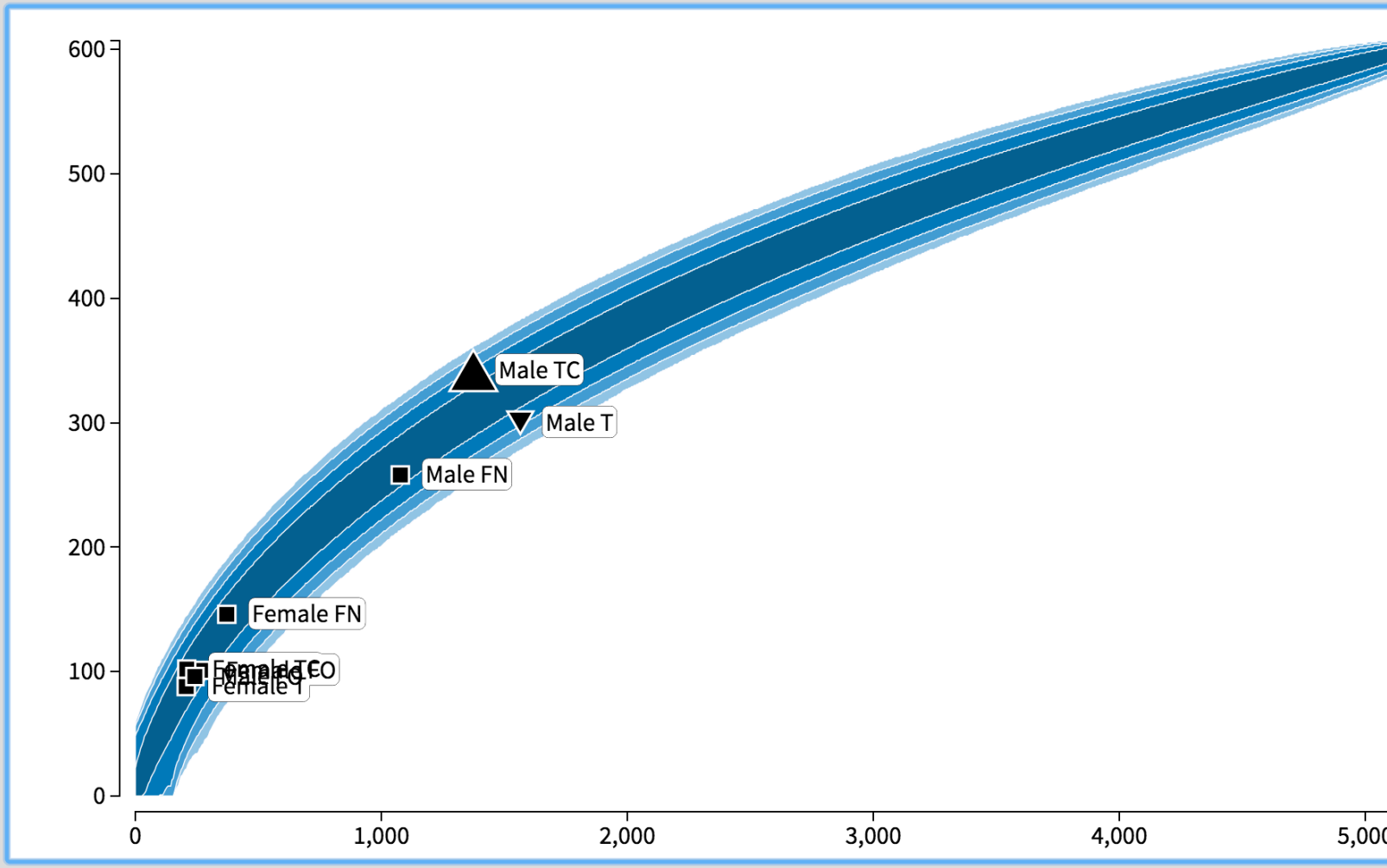
- all
- gender
- **gender + relcode**
- period
- rank
- relcode
- none

Collection

- none
- Female FN
- Female FO
- Female T
- Female TC
- Male FN
- Male FO
- Male T
- **Male TC**

Statistics

- types / running words
- **types / tokens**



The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or** contains 249 hapaxes, 607 types, and 5,318 tokens.

The collection **Male TC** contains 1,374 tokens and 342 types.

Only **1.2%** of random collections with 1,374 tokens contain at least 342 types.

This finding is probably not interesting: the false discovery rate is larger than 0.1.

types2

- Overview
- **Plot**
- Types
- Samples
- Help

Corpus

- **ceec-1680-1800**

Dataset

- **er+or**
- er+or person
- er+or thing

Group

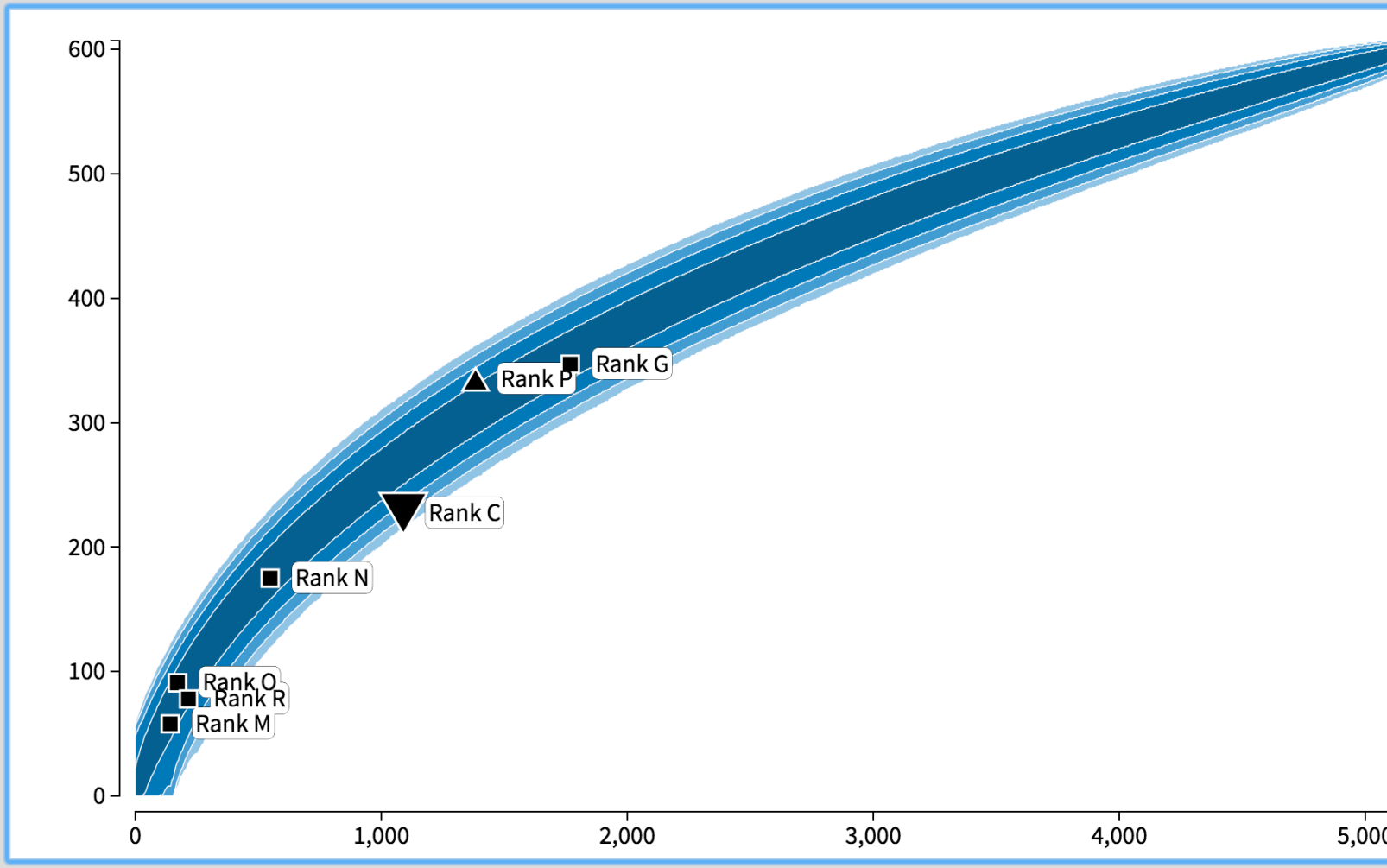
- all
- gender
- gender + relcode
- period
- **rank**
- relcode
- none

Collection

- none
- **Rank C**
- Rank G
- Rank M
- Rank N
- Rank O
- Rank P
- Rank R

Statistics

- types / running words
- **types / tokens**



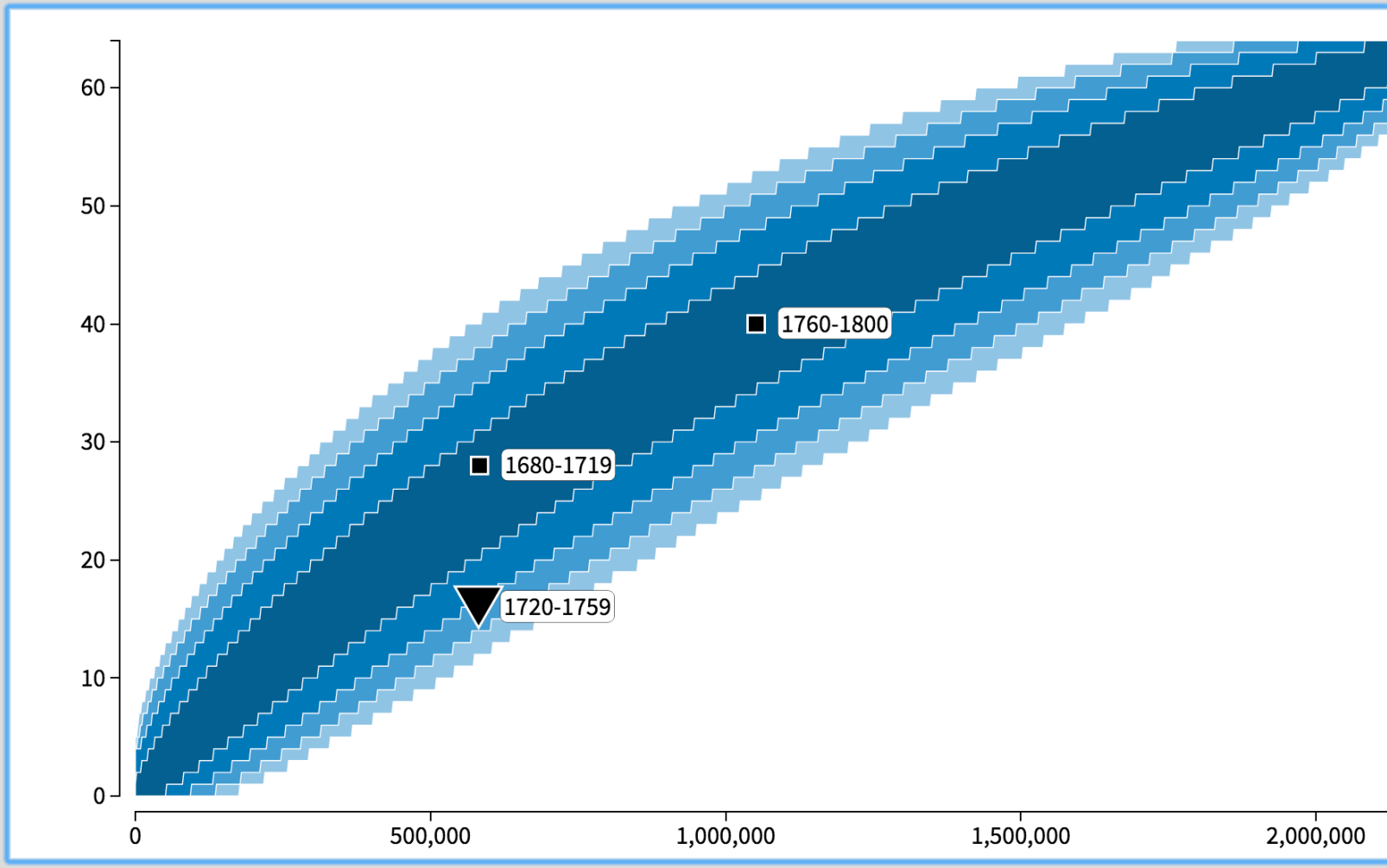
The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or** contains 249 hapaxes, 607 types, and 5,318 tokens.

The collection **Rank C** contains 1,090 tokens and 227 types.

Only **0.18%** of random collections with 1,090 tokens contain at most 227 types.

This finding is probably interesting: the false discovery rate is **0.043**.



The corpus **ceec-1680-1800** contains 579 samples and 2,216,119 running words.

The dataset **er+or thing** contains 35 hapaxes, 64 types, and 167 tokens.

The collection **1720-1759** contains 581,469 running words and 16 types.

Only **0.96%** of random collections with 581,469 running words contain at most 16 types.

This finding is probably not interesting: the false discovery rate is larger than 0.1.

CEEC: Summary of results

- Productivity of -*er* increases over time
  - Inanimate -*er* very infrequent compared to BNC
- Men writing to their close friends overuse -*er*
- Clergy underuse -*er*

Men's use of -*er* to close friends

type	tokens	in collection	fraction	score	samples	in
associator	2	2	1.000		2	
carver	2	2	1.000		2	
cofferer	3	3	1.000		2	
improver	3	3	1.000		2	
miner	2	2	1.000		2	
precentor	2	2	1.000		2	
rider	2	2	1.000		2	
stroller	2	2	1.000		2	
swaggerer	2	2	1.000		2	
worshipper	5	4	0.800		5	
conjurer	10	8	0.800		8	
adorer	1	1	1.000		1	
believer	1	1	1.000		1	

Who are the male users (TC)?

sample	description	running words	tokens
JEVELYN 1680 TC	M John Evelyn	19,594	83
J1WEDGWOOD 1760 TC	M J. Sr Wedgwood	12,676	41
HLIDDELL 1700 TC	M Henry Liddell	36,816	138
TTWINING 1760 TC	M Thomas Twining	11,876	46
WCOWPER 1780 TC	M William Cowper	29,609	76
TTWINING 1780 TC	M Thomas Twining	14,450	30
CLENNOX 1720 TC	M Charles Lennox	6,945	18
HPRIDEAU 1680 TC	M H. Prideaux	12,735	45

CEEC: Interpretation of results

- Productivity of -*er* increases over time
  - Stylistic change or continued semantic expansion? (Säily forthcoming, -*ity*; Dalton-Puffer 1994)
  - Inanimate -*er* very infrequent compared to BNC
    - Later technological developments?
- Men writing to their close friends overuse -*er*
  - Less stable relationship a trigger for productivity? (cf. Wolfson 1990; Säily forthcoming, -*ity*)
- Clergy underuse -*er*
  - ???

Future work

- Further classification
  - Agent/instrument/location?
  - Occasional/habitual/professional agent? (cf. Dalton-Puffer 1994)
  - By word class & etymology of base
- Study both derivation and **inflection**
  - Next: inflectional comparative -*er*
  - Similar variation & change in productivity? → both contribute to syntheticity (Danchev 1992)

References

- Baayen, R.H. 2009. Corpus linguistics in morphology: Morphological productivity. A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics: An International Handbook*, 899–919. Berlin: Mouton de Gruyter.
- Bauer, L. 2001. *Morphological Productivity*. Cambridge: CUP.
- Bauer, L., R. Lieber & I. Plag 2013. *The Oxford Reference Guide to English Morphology*. Oxford: OUP.
- BNC = The *British National Corpus*, version 3 (BNC XML edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.
- CEEC = *Corpora of Early English Correspondence*. Compiled by T. Nevalainen, H. Raumolin-Brunberg et al. at the University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
- Dalton-Puffer, C. 1994. Are Shakespeare's agent nouns different from Chaucer's? – On the dynamics of a derivational sub-system. Kastovsky, D. (ed.), *Studies in Early Modern English*, 45–58. Berlin: Mouton de Gruyter.
- Danchev, A. 1992. Analytic and synthetic developments in English. M. Rissanen et al. (eds.), *History of Englishes*, 25–41. Berlin: Mouton de Gruyter.

References

- Keune, K., R. van Hout & R.H. Baayen. 2006. Socio-geographic variation in morphological productivity in spoken Dutch: A comparison of statistical techniques. J-M. Viprey & Lexicometrica (eds.), *Actes de JADT 2006 : 8es journées internationales d'analyse statistique des données textuelles*, 571–581. Besançon: Université de Franche-Comté.
- Keune, K., R. van Hout & R.H. Baayen 2012. Derivational and lexical productivity across written and spoken Dutch. Chapter 4 in K. Keune, *Explaining Register and Sociolinguistic Variation in the Lexicon: Corpus Studies on Dutch*. PhD dissertation, Radboud University Nijmegen.
- Laws, J.V. & C. Ryder 2014a. *MorphoQuantics*. http://morphoquantics.co.uk/
- Laws, J.V. & C. Ryder 2014b. Getting the measure of derivational morphology in adult speech: A corpus analysis using *MorphoQuantics*. *Language Studies Working Papers*: University of Reading, Vol. 6, 3–17.
- Plag, I. 2003. *Word-Formation in English*. Cambridge: CUP.
- Säily, T. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7(1): 119–141.

References

- Säily, T. Forthcoming. Change or variation? Productivity of the suffixes -*ness* and -*ity*. T. Nevalainen, M. Palander-Collin & T. Säily (eds.), *Sociolinguistic Change in 18th-Century English*.
- Säily, T. & J. Suomela 2009. Comparing type counts: The case of women, men and -*ity* in early English letters. A. Renouf & A. Kehoe (eds.), *Corpus Linguistics: Refinements and Reassessments*, 87–109. Amsterdam: Rodopi.
- Štekauer, P., D. Chapman, S. Tomaščíková & Š. Franko. 2005. Word-formation as creativity within productivity constraints: Sociolinguistic evidence. *Onomasiology Online* 6: 1–55.
- Suomela, J. 2014. *types*2: Type and hapax accumulation curves. http://users.ics.aalto.fi/suomela/types2/
- Suomela, J. 2015. bnc-affix: Analysing productivity of affixes with BNC & MorphoQuantics data. github.com/suomela/bnc-affix
- Wolfson, N. 1990. The bulge: A theory of speech behavior and social distance. *Penn Working Papers in Educational Linguistics* 2(1): 55–83.