Language technology in the service of the humanities

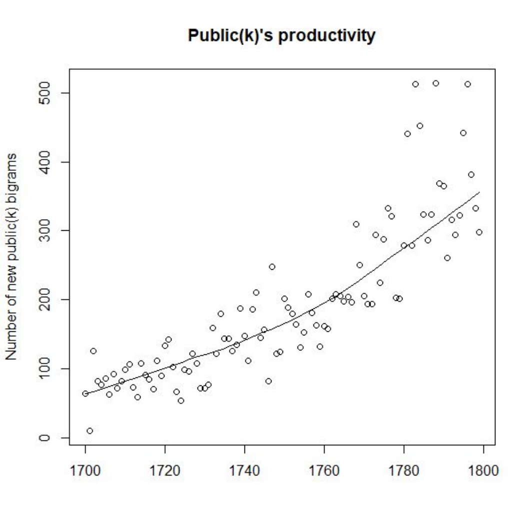
Eetu Mäkelä
Professor (tenure track) in Human Sciences–Computing Interaction
University of Helsinki

Analysing public communication in 18th century Britain and 19th century Finland

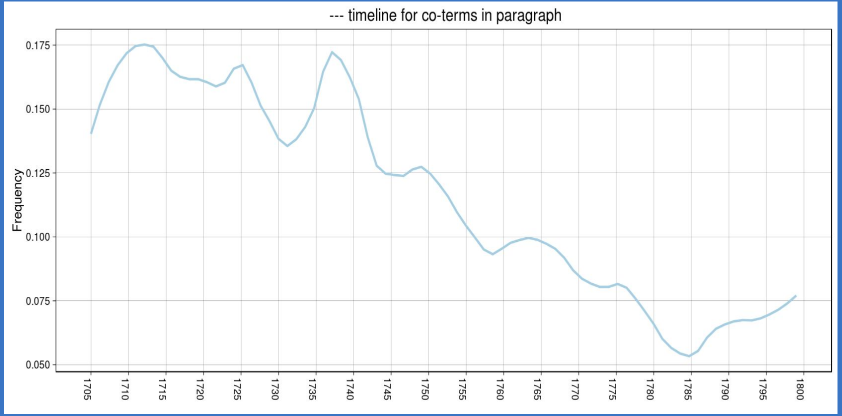
with the COMHIS research group, University of Helsinki

Case public(k)

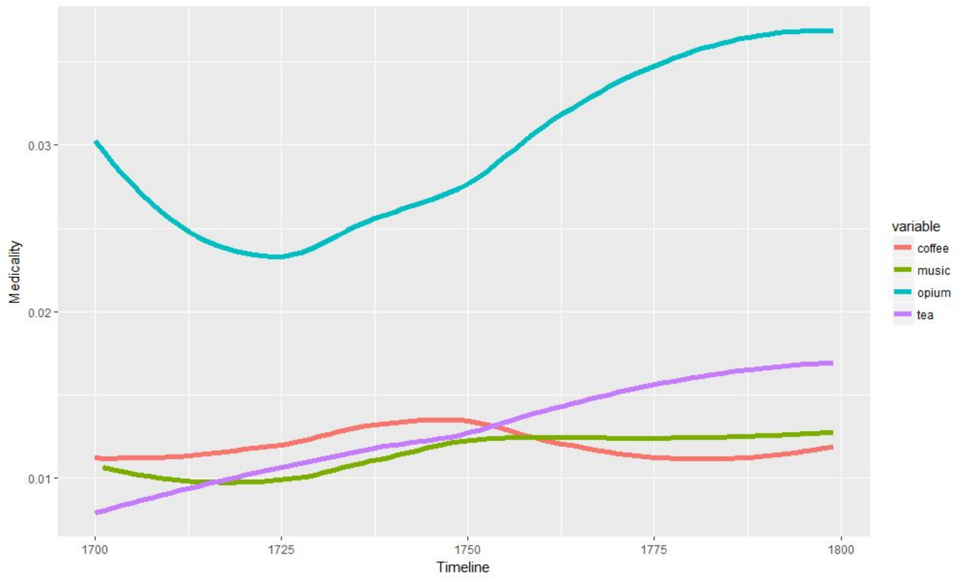
- If a public sphere formed in Britain in the eighteenth century, can we see this in how uses of the word “public” change? How can we historically interpret the changes that we see?

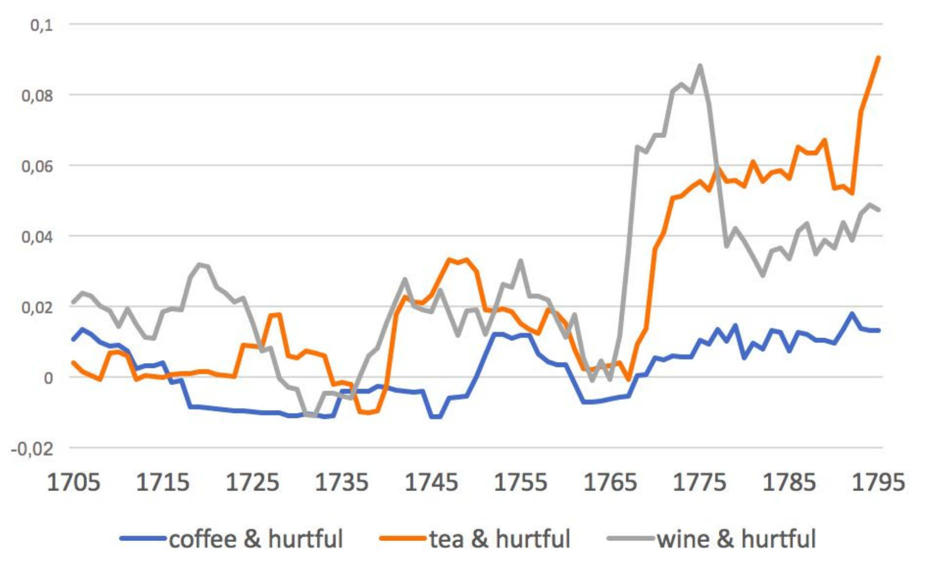


Public + religion



Tracking health aspects of consumables through the 18th Century



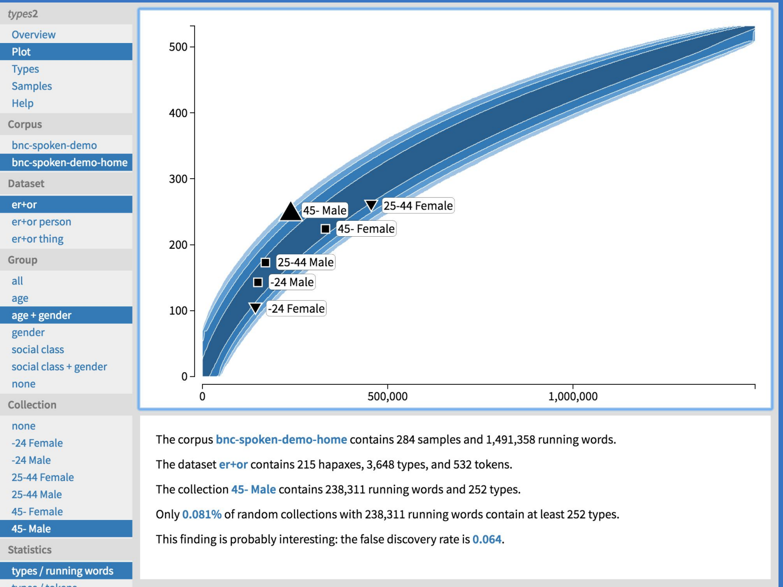


Sociohistorical Language Change

with Tanja Säily and Mika Hämäläinen, University of Helsinki

Questions

- Who creates and adopts new vocabulary in the history of English?
  - Are there differences across social groups?
  - How does new language arise?



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.

Social rank	# of new words
Clergy, lower	7
Genry, lower	7
Gentry, upper	7
Professional	7
Nobility	7
Royalty	6
Clergy, upper	1
Merchant	0
Other non-gentry	0

Etymologies of the new words found

Etymology type	# of new words
Derivative	19
Borrowing	17
Compound	2
Uncertain	1
Zero derivation	1
Borrowing/hybrid	1
Borrowing/zero derivation	1

Etymon language	# of new words
English	22
Latin	11
undetermined	2
Italian	2
French	2
German	1
Dutch	1

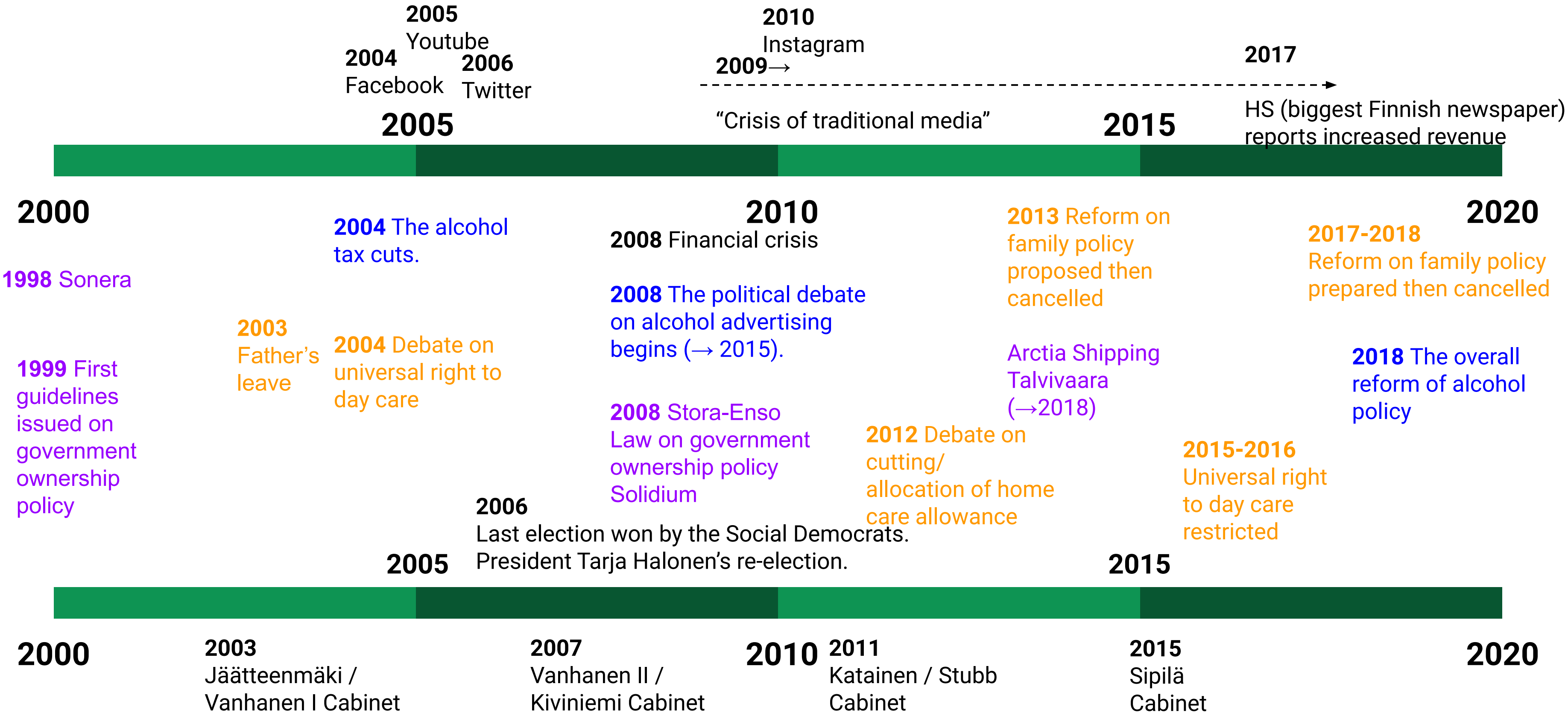
FLOWS OF POWER

media as site and agent of politics

Anu Koivunen (Tampere University)
Eetu Mäkelä (University of Helsinki)

Tampere University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI



2004 Facebook

2005 Youtube

2006 Twitter

2010 Instagram

2009→

“Crisis of traditional media”

2017

HS (biggest Finnish newspaper) reports increased revenue

2000 **2005** **2010** **2015** **2020**

1998 Sonera

1999 First guidelines issued on government ownership policy

2003 Father’s leave

2004 The alcohol tax cuts.

2004 Debate on universal right to day care

2008 Financial crisis

2008 The political debate on alcohol advertising begins (→ 2015).

2008 Stora-Enso Law on government ownership policy Solidium

2006
Last election won by the Social Democrats.
President Tarja Halonen’s re-election.

2012 Debate on cutting/ allocation of home care allowance

2013 Reform on family policy proposed then cancelled

Arctia Shipping Talvivaara (→2018)

2015-2016 Universal right to day care restricted

2017-2018 Reform on family policy prepared then cancelled

2018 The overall reform of alcohol policy

2000 **2005** **2010** **2015** **2020**

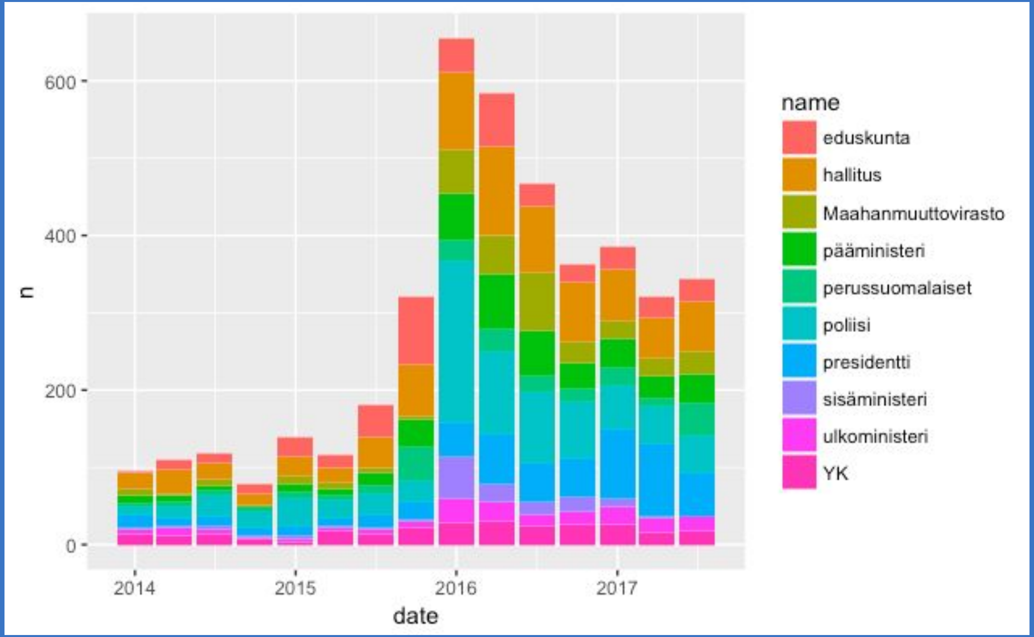
2003 Jäätteenmäki / Vanhanen I Cabinet

2007 Vanhanen II / Kiviniemi Cabinet

2011 Katainen / Stubb Cabinet

2015 Sipilä Cabinet

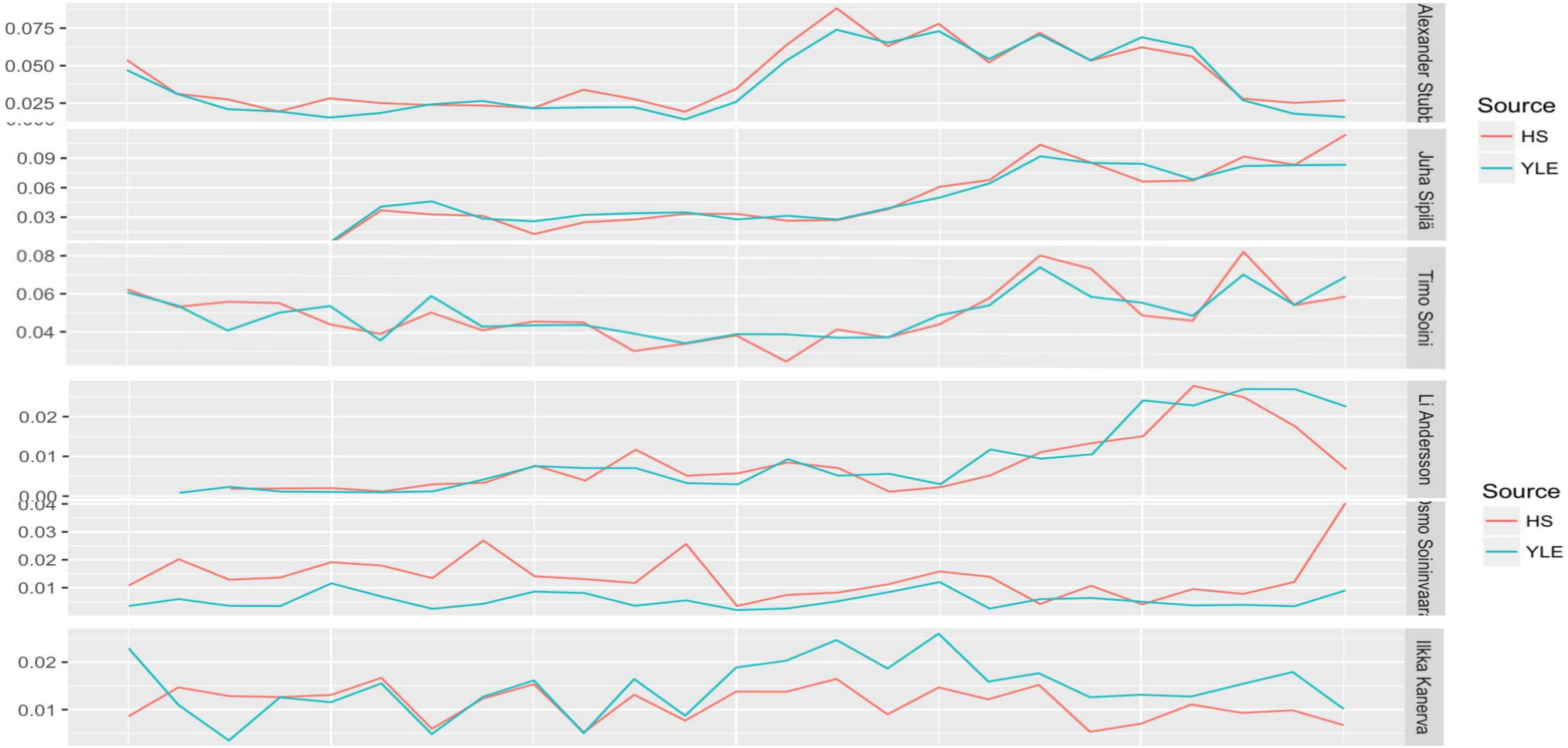
Authorities referred to in the debate on immigration in YLE by year



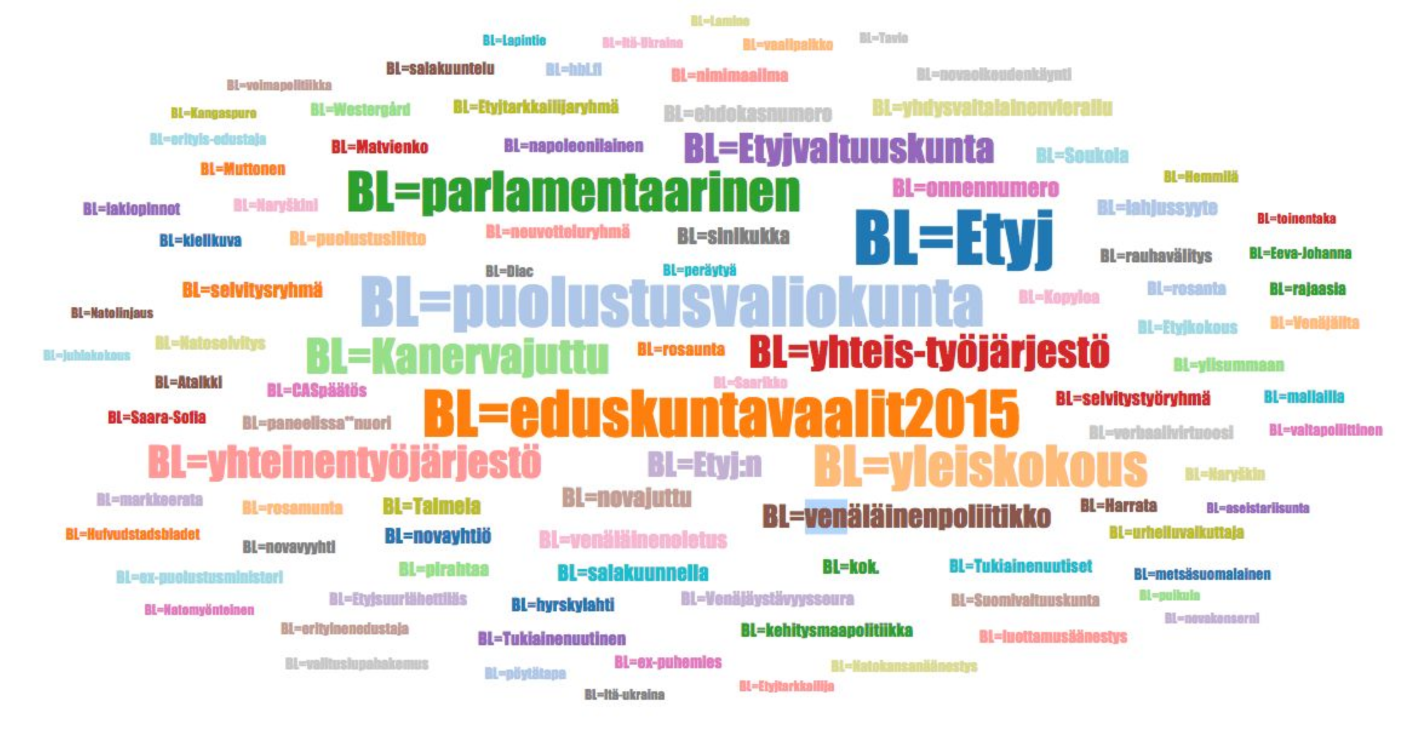
UNIVERSITY OF HELSINKI
FACULTY OF ARTS

HELDIG
Helsinki Centre for Digital Humanities

Media trajectories of prominent Finnish politicians based on proportion of mentions



Words associated with Ilkka Kanerva in YLE



Groups of people often appearing together in association with Ilkka Kanerva in YLE

- Matti Louekoski, Matti Vanhala, Sinikka Salo
- Arto Merisalo, Jorma Kalske, Tapani Yli-Saunamäki, Toivo Sukari
- Ari Ruotsalainen, Arto Merisalo, Kyösti Kakkonen, Tapani Yli-Saunamäki, Toivo Sukari
- Anne-Mari Virolainen, Esko Kiviranta, Heli Paasio, Ilkka Kantola, Katja Taimela, Pertti Hemmilä, Petteri Orpo, Stefan Wallin, Ville Niinistö
- Anne-Mari Virolainen, Annika Lapintie, Esko Kiviranta, Heli Paasio, Ilkka Kantola, Katja Taimela, Pertti Hemmilä, Petteri Orpo, Stefan Wallin, Ville Niinistö

Adjectives associated with Juha Sipilä in YLE

- suuri
- uusi
- vahva
- hurja
- hyvä
- rehti
- avoin
- rauhallinen
- tervetullut
- lupsakka
- uskonnollinen
- raikas poikkeus
- nopea

“Kirjoittaja vertaa Juha Sipilän taivalta heti alussa Julius Caesarin rakettimaiseen voittokulkuun.”

Intertextuality and thematic networks in Finnic regional poetic cultures

with Kati Kallio, Finnish Literature Society

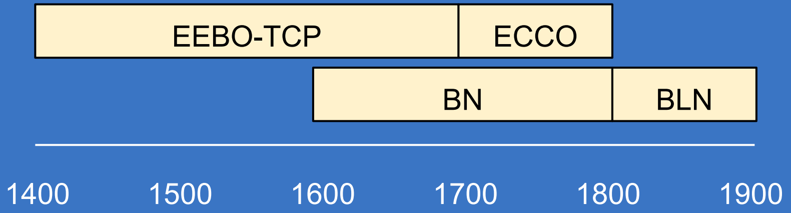
Research questions

- Tracing multiple levels of commonalities across varying Finnish and Estonian corpora:
  - Characters / individual words
  - Verses
  - Formulae (small repeating bits, e.g. sirkun siivet/linnun siivet/tedre tiivad)
  - Themes

Needs from language technology

Materials

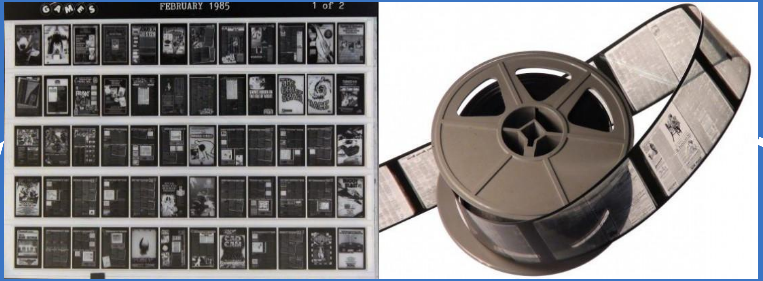
- Eighteenth Century Collections Online: ~210,000 books (~84% of those known to exist), ~10 billion running words 1700–1800 (ECCO)
- Early English Books Online: ~60,000 books (~53% of those known to exist), ~1.4 billion running words 1400–1700 (EEBO-TCP)
- British Library Newspapers: ~655,000 newspaper issues (uneven selection), ~40 billion running words 1800–1900 (BLN)
- Burney & Nichols Collections: ~230,000 pamphlets and newspapers, ~2.7 billion running words before 1600–1800 (BN)



EEBO-TCP
ECCO
BN
BLN

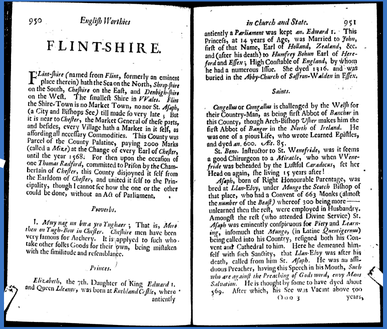
Data production pipelines: Early English Books Online

Physical books



Microfilms

Electronic Image Scans



950 *Engliſh Worthies*

FLINT-SHIRE.

Flint-ſhire (named from *Flint*, formerly an eminent place therein) hath the Sea on the North, *Shrop-ſhire* on the South, *Cheſhire* on the Eaſt, and *Denbigh-ſhire* on the Weſt. The ſmalleſt Shire in *VVales*. *Flint* the Shire-Town is no Market Town, no nor St. *Aſaph*, (a City and Biſhops See) till made ſo very late; But it is near to *Cheſter*, the Market General of theſe parts, and beſides, every Village hath a Market in it ſelf, as affording all neceſſary Commodities. This County was Parcel of the County Palatine, paying 2000 Marks (called a *Mize*) at the Change of every Earl of *Cheſter*, until the year 1568. For then upon the occaſion of one *Thomas Radford*, committed to Priſon by the Chamberlain of *Cheſter*, this County disjoyned it ſelf from the Earldom of *Cheſter*, and united it ſelf to the Principality, though I cannot ſee how the one or the other could be done, without an Act of Parliament.

Proverbs.

1. *Mwy nag un bwa yro Tnghaer*; That is, *More then on Tugh-Bow in Cheſter*. *Cheſhire* men have been very famous for Archery. It is applyed to ſuch who take other folks Goods for their own, being miſtaken with the ſimilitude and reſemblance.

Princes.

Elizabeth, the 7th. Daughter of King *Edward* 1. and Queen *Lleanor*, was born at *Ruthland Caſtle*, where antiently a *Parliament* was kept *an. Edward* 1. This Princeſs, at 14 years of Age, was Married to *John*, firſt of that Name, Earl of *Holland*, *Zealand*, &c. and (after his death) to *Humfrey Bohun* Earl of *Hereford* and *Eſſex*; High Conſtable of *England*, by whom he had a numerous Iſſue. She dyed 1316. and was buried in the *Abby-Church* of *Saffron-Walden* in *Eſſex*.

in Church and State. 951

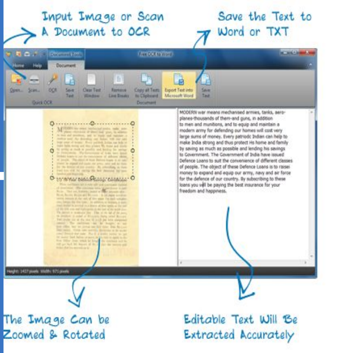
Saints.

Congellus or *Comgallus* is challenged by the *Welſh* for their Country-Man, as being firſt Abbot of *Banchor* in this County, though Arch-Biſhop *Uſher* makes him the firſt Abbot of *Bangor* in the *North* of *Ireland*. He was one of a pious Life, who wrote Learned Epiſtles, and dyed *an*. 600. *Æt*. 85.

St. *Beno*. Inſtructor to St. *Wenefride*, was it ſeems a good Chirurgeon to a *Miracle*, who when *VVenefride* was beheaded by the Luſtful *Caradocus*, ſet her Head on again, ſhe living 15 years after!

Aſaph, born of Right Honourable Parentage, was bred at *Llan-Elvy*, under *Mungo* the *Scotch* Biſhop of that place, who had a Convent of 663 Monks (almoſt the *number* of the *Beaſt*) whereof 300 being more—unlearned then the reſt, were employed in Husbandry. Amongſt the reſt (who attended Divine Service) St. *Aſaph* was eminently conſpicuous for *Piety* and *Learning*, inſomuch that *Mungo*, (in Latine *Quentigernus*) being called into his Country, reſigned both his Convent and Cathedral to him. Here he demeaned himſelf with ſuch Sanctity, that *Llan-Elvy* was after his death, called from him St. *Aſaph*. He was an aſſiduous Preacher, having this Speech in his Mouth, *Such who are againſt the Preaching of Gods word, envy Mans Salvation*. He is thought by ſome to have dyed about 569. After which, his See was Vacant above 500 years,

Ooo 3



Input Image or Scan A Document to OCR

Save the Text to Word or TXT

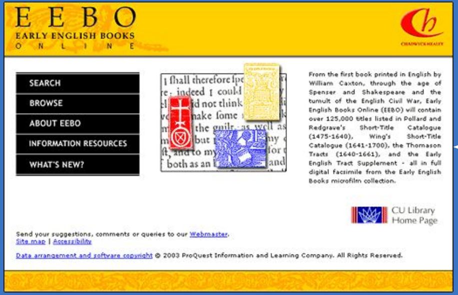
MODERN war means mechanised armies, tanks, aeroplanes-thousands of them-and guns, in addition to men and munitions, and to equip and maintain a modern army for defending our homes will cost very large sums of money. Every patriotic Indian can help to make India strong and thus protect his home and family by saving as much as possible and lending his savings to Government. The Government of India have issued Defence Loans to suit the convenience of different classes of people. The object of these Defence Loans is to raise money to expand and equip our army, navy and air force for the defence of our country. By subscribing to these loans you will be paying the best insurance for your freedom and happiness.

Height: 1437 pixels Width: 971 pixels

The Image Can be Zoomed & Rotated

Editable Text Will Be Extracted Accurately

OCR



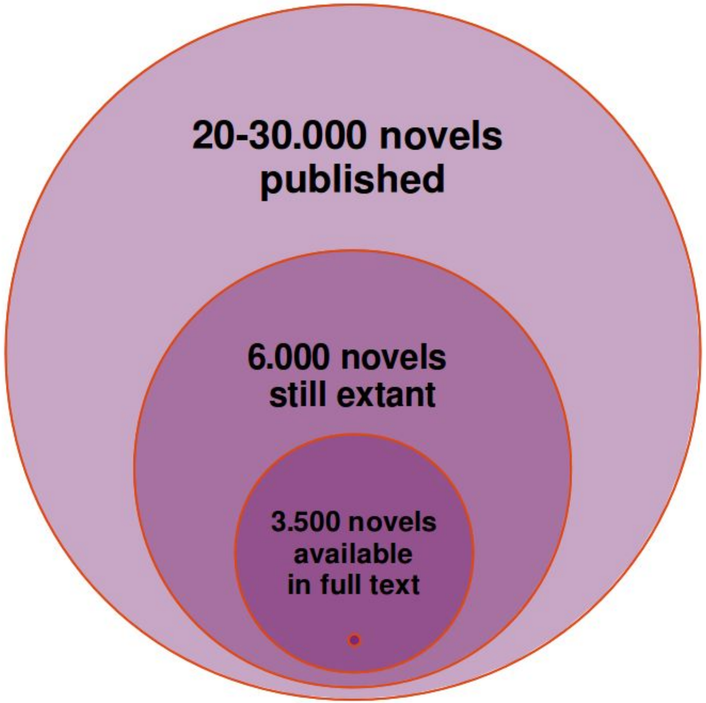
From the first book printed in English by William Caxton, through the age of Spenser and Shakespeare and the tumult of the English Civil War, Early English Books Online (EEBO) will contain over 125,000 titles listed in Pollard and Redgrave's Short-Title Catalogue (1475-1640), Wing's Short-Title Catalogue (1641-1700), the Thomason Tracts (1640-1661), and the Early English Tract Supplement - all in full digital facsimile from the Early English Books microfilm collection.

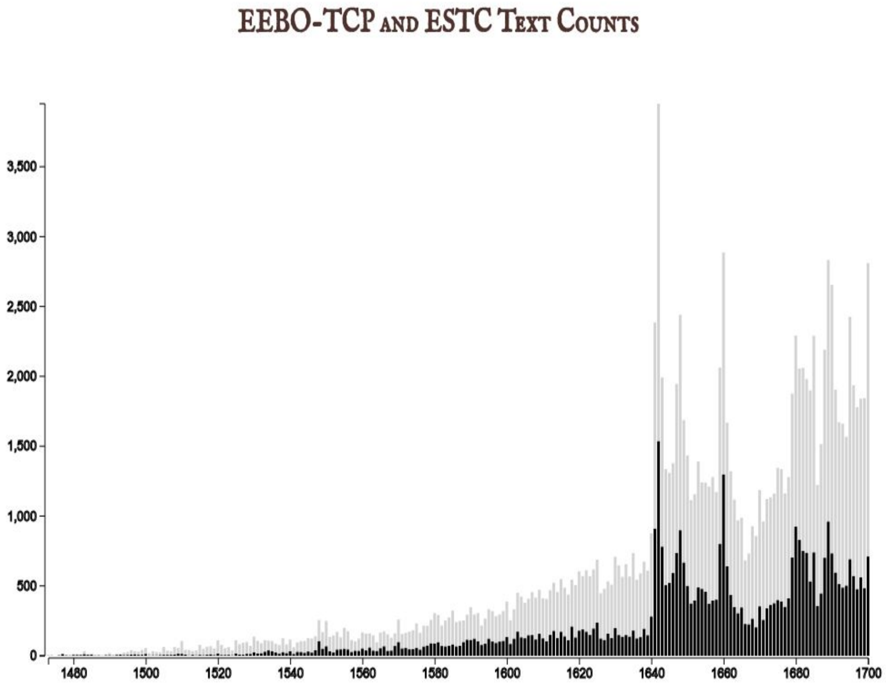
Send your suggestions, comments or queries to our Webmaster.
Site map | Accessibility

Data arrangement and software copyright © 2003 ProQuest Information and Learning Company. All Rights Reserved.

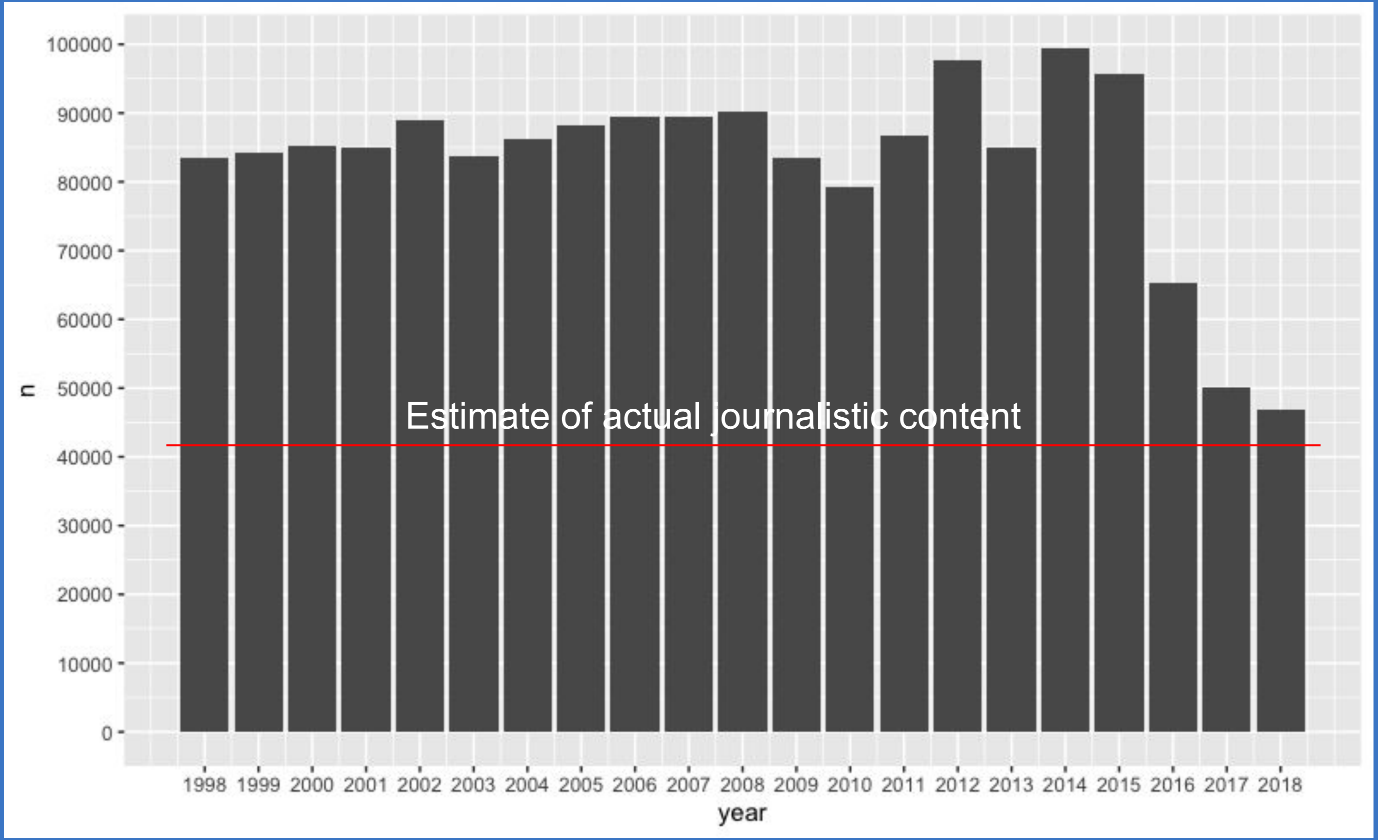
Early English Books Online

But what's in there, e.g. genre-wise?


20-30.000 novels published
6.000 novels still extant
3.500 novels available in full text

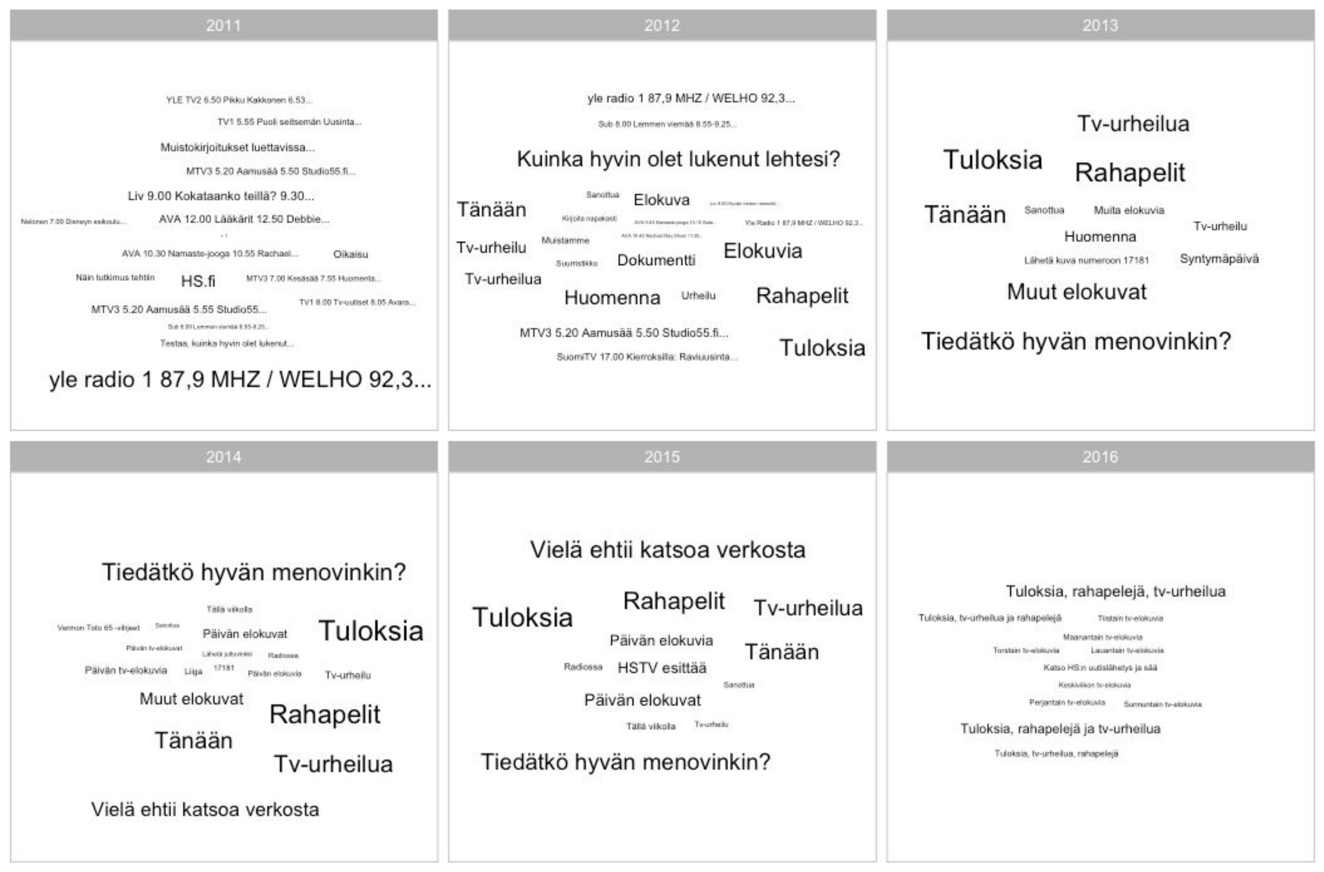

EEBO-TCP and ESTC Text Counts

Number of articles per year in HS

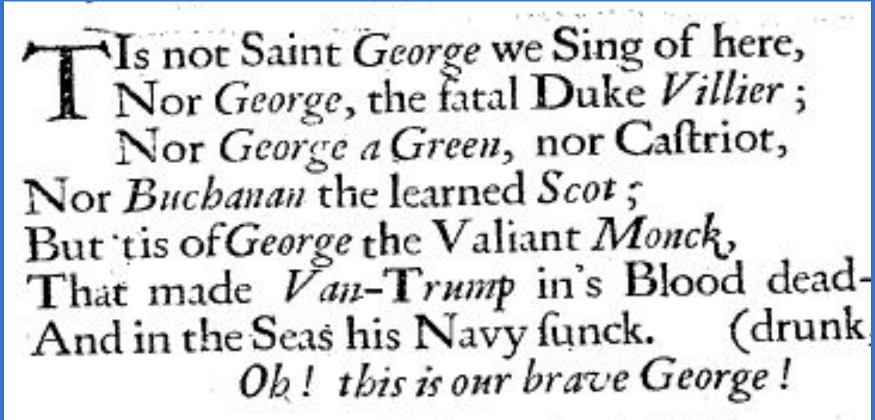


UNIVERSITY OF HELSINKI
FACULTY OF ARTS

Surfacing distributional anomalies

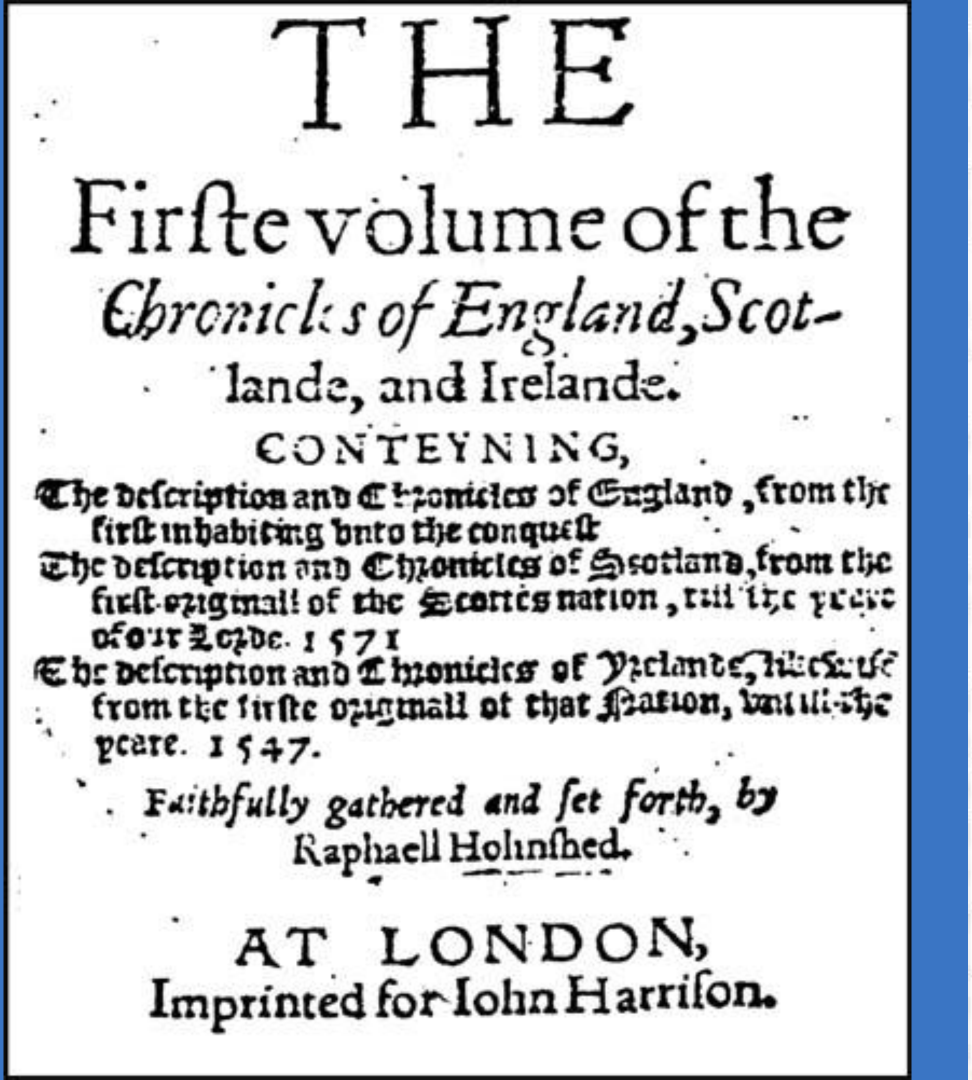


Relatively good OCR



TIs not Saint *George* we Sing of here,
Nor *George*, the fatal Duke *Villier* ;
Nor *George a Green*, nor Caſtriot,
Nor *Buchanan* the learned *Scot* ;
But 'tis of *George* the Valiant *Monck*,
That made *Van-Trump* in's Blood dead-
And in the Seas his Navy ſunck. (drunk,
Oh! this is our brave George !

Isw-not- Saint George we Sing of here,
Nor George, the fatal Duke Villier ;
Nor George a Green, nor Castriot,
Nor Buchanan the learned S cot q
But us of George the Valiant Monck,
That made Van-Trump in'S Blood deod
and in theseus his Navy snuck. (drunk,
Ok I this is our brave George !



THE

Firſte volume of the
*Chronicles of England, Scot-
lande, and Irelande.*

CONTEYNING,

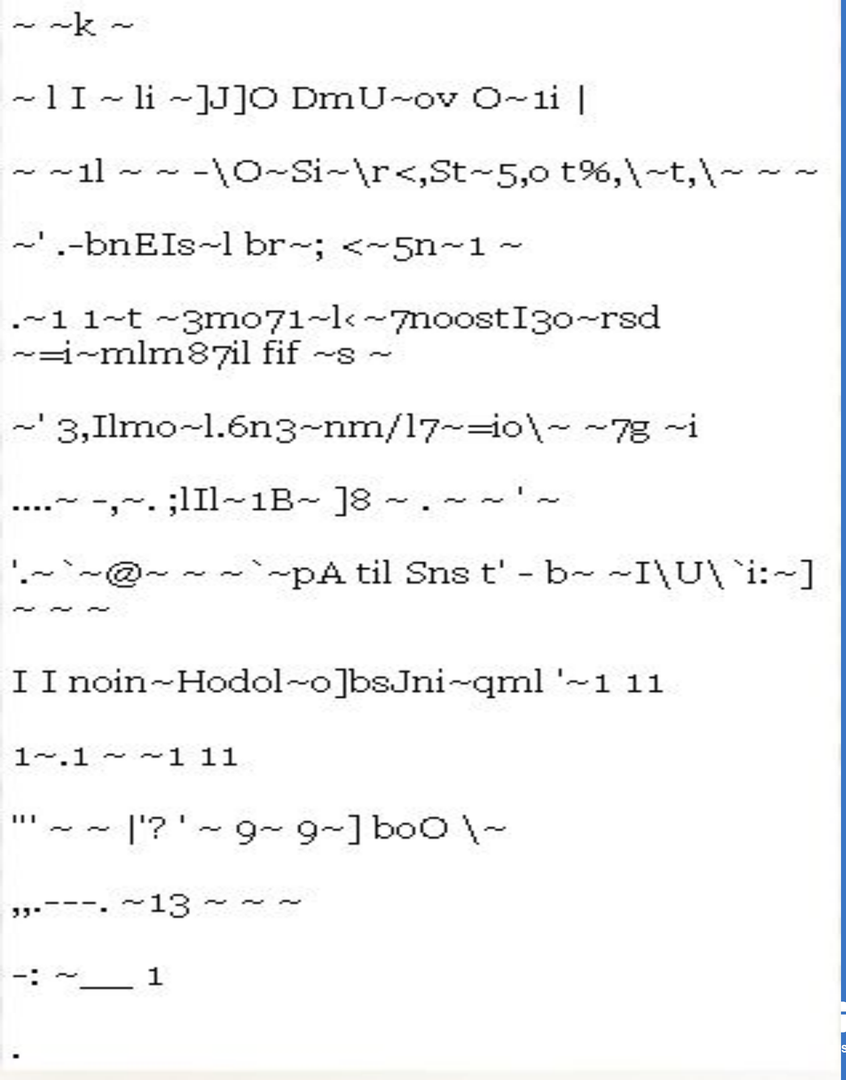
The deſcription and Chronicles of England, from the firſt inhabiting vnto the conqueſt

The deſcription and Chronicles of Scotland, from the firſt originall of the Scottes nation, till the yeare of our Lorde. 1571

The deſcription and Chronicles of Yrelande, likewiſe from the firſte originall of that Nation, vntill the yeare. 1547.

Faithfully gathered and ſet forth, by
Raphaell Holinſhed.

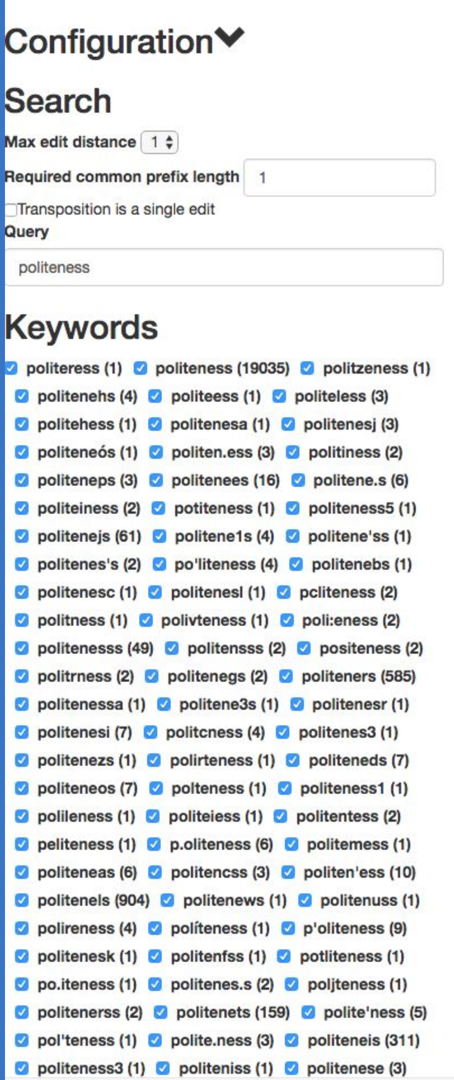
AT LONDON,
Imprinted for Iohn Harriſon.



```
~ ~k ~

~ l I ~ li ~]J]O DmU~ov O~1i |

~ ~1l ~ ~ -\O~Si~\r<,St~5,o t%,\~t,\~ ~ ~

~' .-bnEIs~l br~; <~5n~1 ~

.~1 1~t ~3mo71~l‹ ~7noostI3o~rsd
~=i~mlm87il fif ~s ~

~' 3,Ilmo~l.6n3~nm/l7~=io\~ ~7g ~i

....~ -,~. ;lIl~1B~ ]8 ~ . ~ ~ ' ~

'.~ `~@~ ~ ~ `~pA til Sns t' - b~ ~I\U\ `i:~]
~ ~ ~

I I noin~Hodol~o]bsJni~qml '~1 11

1~.1 ~ ~1 11

''' ~ ~ |'? ' ~ 9~ 9~] boO \~

,,.---. ~13 ~ ~ ~

-: ~___ 1

.
```

OCR problems

- vectors.most_similar("king") → [('k~ing', 0.88), ('kiing', 0.86), ('"king"', 0.84), ('emperor', 0.823), ('icing', 0.83), ('kring', 0.82), ('kcing', 0.82), ('.king', 0.82), ('kting', 0.81), ('iking', 0.81)]



Configuration❯

Search

Max edit distance 1

Required common prefix length 1

□ Transposition is a single edit

Query

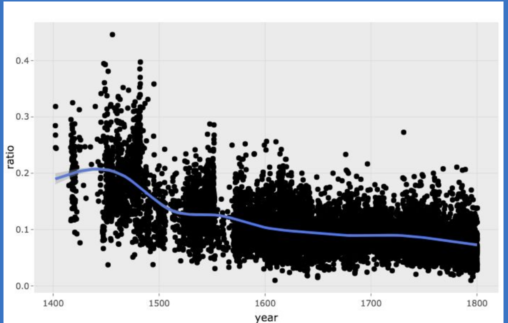
politeness

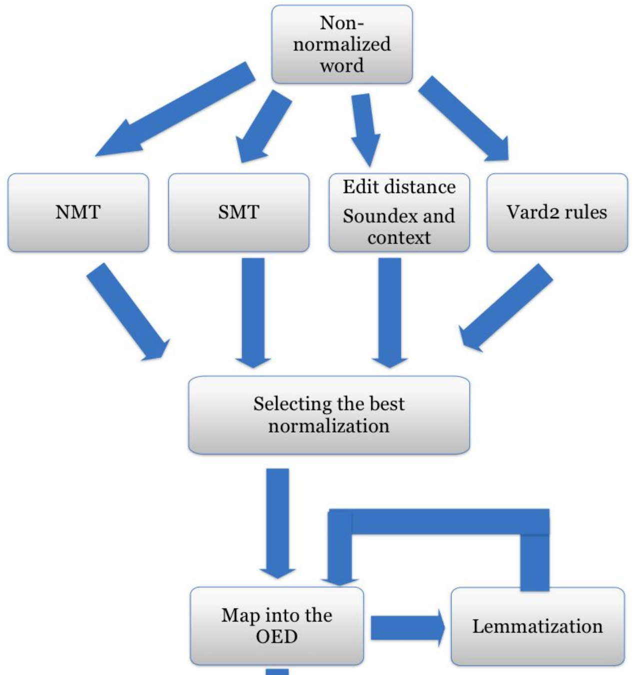
Keywords

☑ politeress (1) ☑ politeness (19035) ☑ politzeness (1)
☑ politenehs (4) ☑ politeess (1) ☑ politeless (3)
☑ politehess (1) ☑ politenesa (1) ☑ politenesj (3)
☑ politeneós (1) ☑ politen.ess (3) ☑ politiness (2)
☑ politeneps (3) ☑ politenees (16) ☑ politene.s (6)
☑ politeiness (2) ☑ potiteness (1) ☑ politeness5 (1)
☑ politenejs (61) ☑ politene1s (4) ☑ politene'ss (1)
☑ politenes's (2) ☑ po'liteness (4) ☑ politenebs (1)
☑ politenesc (1) ☑ politenesl (1) ☑ pcliteness (2)
☑ politness (1) ☑ polivteness (1) ☑ poli:eness (2)
☑ politenesss (49) ☑ politensss (2) ☑ positeness (2)
☑ politrness (2) ☑ politenegs (2) ☑ politeners (585)
☑ politenessa (1) ☑ politene3s (1) ☑ politenesr (1)
☑ politenesi (7) ☑ politcness (4) ☑ politenes3 (1)
☑ politenezs (1) ☑ polirteness (1) ☑ politeneds (7)
☑ politeneos (7) ☑ polteness (1) ☑ politeness1 (1)
☑ polileness (1) ☑ politeiess (1) ☑ politentess (2)
☑ peliteness (1) ☑ p.oliteness (6) ☑ politemess (1)
☑ politeneas (6) ☑ politencss (3) ☑ politen'ess (10)
☑ politenels (904) ☑ politenews (1) ☑ politenuss (1)
☑ polireness (4) ☑ polĺteness (1) ☑ p'oliteness (9)
☑ politenesk (1) ☑ politenfss (1) ☑ potliteness (1)
☑ po.iteness (1) ☑ politenes.s (2) ☑ poljteness (1)
☑ politenerss (2) ☑ politenets (159) ☑ polite'ness (5)
☑ pol'teness (1) ☑ polite.ness (3) ☑ politeneis (311)
☑ politeness3 (1) ☑ politeniss (1) ☑ politenese (3)

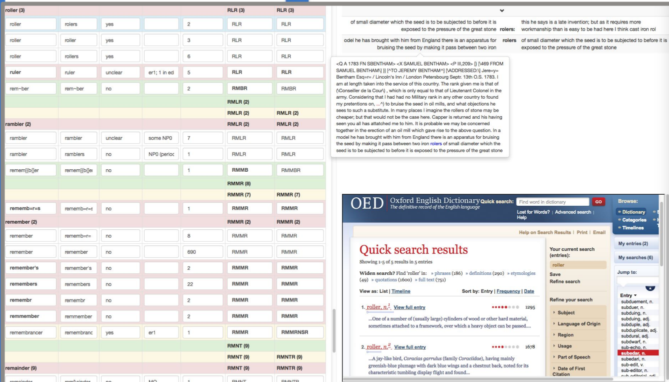
Spelling variation

- “... Ships as shall be in readynesse for that service, and this matter requireing the greatest secrecy,... Your very affectionett freind” (Queen Anne, 1704)
- “Right worshipfull and my most entierly beloude moder, in the most louly maner I recommaund me vnto youre gode moderhode,” (Elizabeth Poynings, 1459)





Non-normalized word
NMT
SMT
Edit distance Soundex and context
Vard2 rules
Selecting the best normalization
Map into the OED
Lemmatization



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)

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

OED Oxford English Dictionary
The definitive record of the English language

Quick search results

Showing 1-5 of 5 results in 5 entries

Widen search? Find ‘roller’ in: » phrases (186) » definitions (290) » etymologies (49) » quotations (1600) » full text (751)

View as: List | Timeline Sort by: Entry | Frequency | Date

1. roller, n.1 View full entry ●●●●●○○○ 1295
...One of a number of (usually large) cylinders of wood or other hard material, sometimes attached to a framework, over which a heavy object can be passed....

2. roller, n.2 View full entry ●●●●○○○○ 1678
...A jay-like bird, *Coracias garrulus* (family *Coraciidae*), having mainly greenish-blue plumage with dark blue wings and a chestnut back, noted for its characteristic tumbling display flight and found...

Your current search (entries):
roller
Save
Refine search

Refine your search
- Subject
- Language of Origin
- Region
- Usage
- Part of Speech
- Date of First Citation

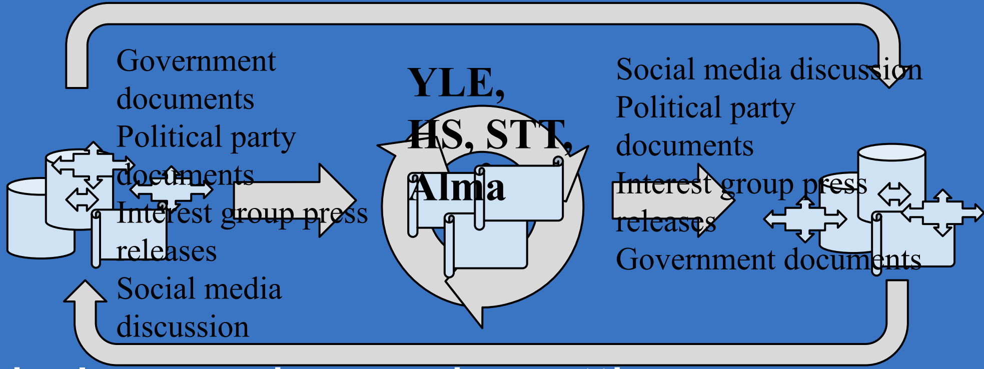
Entry
subduement, n.
subduer, n.
subduing, n.
subduing, adj.
subduple, adj.
subduplicate, adj.
subdural, adj.
subdwarf, n.
sub-echo, n.
subedar, n.
subedari, n.
sub-edit, v.
sub-editor, n.

Linguistic complexity

- “hopia* vaski*”~10 (any word starting with hopia within 10 words of any word starting with vaski)
- (/pi[sk]ku.*/ OR pien*) AND /pi[ij]at.*/
- |infl_fi_~2:Väinämöinen (any [dialectal] variation at max two edit distances from any inflected form of the Finnish name Väinämöinen)

Focused, versatile, configurable NLP

- Interaction between the computational and the qualitative
- Operationalising indicators to measure the framing of different actors:
  - Tone
  - Standing
  - Polarisation
  - Affect
  - Fact/opinion -balance
  - Narrativity
- Tracking of the flow of vocabulary and agenda-setting power between actors and the media


Government documents
Political party documents
Interest group press releases
Social media discussion
YLE, HS, STT, Alma
Social media discussion
Political party documents
Interest group press releases
Government documents

Needs from language technology

- Quantifying noise and bias (OCR errors, genre balance, variation, …)
    - Quantifying the amount of OCR errors/variation in different parts of a collection
    - Unsupervised genre identification → configurable genre classification
- Methods to clean noise and bias
    - OCR post-correction, better OCR through better language models?
    - Subsampling based on genre information
- Noise/variation-resistant NLP
    - Noise-resistant embeddings
    - Flexible variant normalisation, variation-sensitive querying
- Focused, configurable NLP
    - E.g. from sentiment analysis to the identification of hedging, tone, standing, polarisation, fact/opinion -balance, narrativity, …
    - + everything else you saw before

eetu.makela@helsinki.fi
http://iki.fi/eetu.makela

This presentation:
http://j.mp/lt-hums