

TEXT REUSE IN DH RESEARCH

Matteo Romanello (UNIL)

Text Reuse at Scale workshop – University of Luxembourg, 23-24 November 2022

AJMC

AJAX
MULTI
COMMENTARY

Unil

UNIL | Université de Lausanne

SUMMARY

1. Definition, applications, methods
2. Text reuse in *impresso*
3. Text reuse *beyond* *impresso*
4. Text reuse at scale

DEFINITION IN CONTEXT

What do we mean by *text reuse detection* and when is it used?

What research questions are investigated by means of it?

DEFINITION

Text Reuse (TR)

“The meaningful reiteration of text, usually beyond the simple repetition of common language.”

(Romanello et al. 2014)

Text reuse detection (TRD)

The task of automatically detecting reiterations (reuses) of text, usually carried out in an unsupervised setting.

CONTEXTS

“*meaningful* reiteration”: function of domain, document type and research question/application.

Scholarly publications

Inf...
Matthias Döring, Joachim Bück, Georg Friedrich, Alejandro Pironti et al. "geno2...
Otherwise, the annotated species of the reference sequence with the greatest alignment score is used. An alignment is considered a high-similarity alignment if it satisfies two similarity criteria, which are defined by dividing the number of matching amino acids in the alignment either by the length of the alignment (alignment similarity) or by the length of the reference sequence (reference similarity). For HIV-1 sequences, a minimal alignment similarity of 60% and a minimal reference similarity of 50% is used for all regions, except for the reverse transcriptase (RT). Since all major drug resistance mutations are located within the first half of the gene, the RT region is frequently merely partially amplified. Thus, we use a reference sim. length of only 50% for the RT. Due to the regular is, alignments exhibiting a high degree of similarity between query and reference sequence. An alignment is considered a high-similarity alignment if it satisfies two similarity criteria, which are defined by dividing the number of matching amino acids in the alignment either by the length of the alignment (alignment similarity) or by the length of the reference sequence (reference similarity).

(Image [source](#))

Quotations, plagiarism
Academic writing/bibliometrics

Newspapers



Copy/shake & paste, repurposing
Printing culture, virality of news

Literature

Shakespeare, William (1564-1616)
1710 - Macbeth

MACBETH 29
Macb. Thus comes my Fit again; I had it before perfect,
From a Father deadlier than that
The world did ever see: Proceed, Air;
See now the child I wish every Usurper and Peer;
Macb. Banish me from here, Heaven,
With twenty thousand swords upon his Head,
The hell which was my Hell!
Macb. There the grand Scepter lies; the Womán that's dead
High Names that in time will Venge behead,
Though in prison it wants Sleep, to move,
To move you that hear further.
La. Macb. My Royal Lord, you spill the Full,
The Scepter to Macb. Macb.
Macb. Enter the Ghost of Banquo, and first a Macbeth's plea.
Macb. Let good Digitation was an Agitate.
And Health to both.
La. Macb. I find your Highness to be.
Macb. Had we but here our Country's Honour;
When the great Justice of our Answer pruned;
When we may fully challenge the Unleashed.
Macb. The Answer, Sir.
La. Macb. Open the portals; think your Highness
To you with your Company.
Macb. I'll be down. The Child's full.
La. Macb. Here is a Place refer to be.
Macb. When, Sir?
La. Macb. When do the move your Highness?
Macb. Which of you have done that?
La. Macb. Dost thou?
Macb. Thus we'll see by I had it; were that
The great Lords at me.
La. Macb. Sir, wretched Frenzy, my Lord is often this,
And look here from his Youth; For long your love,
The fit is your father, if you take notice of it,
You shall understand, and possible his father.
His mother is'll be well again.
Macb. And hold one that dare look on that
As you should!
Macb. O, proper first!
Macb. O, proper first!
Macb. This is the very living of our days
This is the very living of our days
And you to answer. O, think this and Sorrow
(Macbeth's name) would not increase
A Woman's there, rebuked by her Guardian,
Why do you then that with all's done
You look but get a Chair.

Allusions, paraphrases, quotations
Intertextuality, lit. reception

APPLICATIONS

- To identify multiple editions of the same work(s) in a corpus
- To find quotations of a text when target works are known
- To identify/filter out duplicates before further processing
- To study the virality and spread of texts
- ...

Primary text (v. 1-3)

ἌΕΙ μὲν, ὦ παῖ Λαρτίου, δέδορκά σε
πείραν τιν' ἐχθρῶν ἀρπάσαι θηρώμενον·
καὶ νῦν ἐπὶ σκηναῖς σε ναυτικαῖς ὄρω

Lemmata

1 αὖ μὲν, followed in 3 by καὶ νῦν:
cp. *Tr.* 689–691 ἔχρισα μὲν...κᾶθηκα
(n.): *Lucian Dialog. marin.* 8 πάλαι μὲν

Λαρτίου, as in 380; but Λαερτίου in
101, and Λαέρτου in 1393. Λαέρτης is
the only Homeric form (*Ph.* 87 n.), but
Eur., like Soph., uses all three. In

■ Line number anchor

■ Word anchor

■ Reference to primary literature

■ Reference to tutor text

METHODS

Newspapers domain

Challenges:

- OCR noise
- Scale of corpora

Methods:

- Character-level alignment → BLAST (Vesanto et al. 2017)
- Ngram-level alignment → Passim (Smith et al. 2015)

Literature domain

Challenges:

- Quotations, allusions, paraphrases
- Need to go beyond repetition of n-grams

Methods:

- TRACER toolbox for historical text reuse (Büchler et al. 2014)
- Extend the feature set → syntax, metre, semantic context (Coffee et al. 2012, Scheirer et al. 2016, Moritz & Steding 2018)
- Word embeddings-based similarity (Liebl & Burghardt 2020)

IMPLEMENTATIONS

- The [R textreuse package](#) (R) written by Lincoln Mullen
- [TRACER](#) (Java) developed by Marco Böhler and colleagues
- [Basic Local Alignment Search Tool \(BLAST\)](#)
- [Tesseract](#) (PHP, Perl)
- [TextPAIR \(Pairwise Alignment for Intertextual Relations\)](#)
- [Passim](#) (Scala) developed by David Smith (Northeastern University)

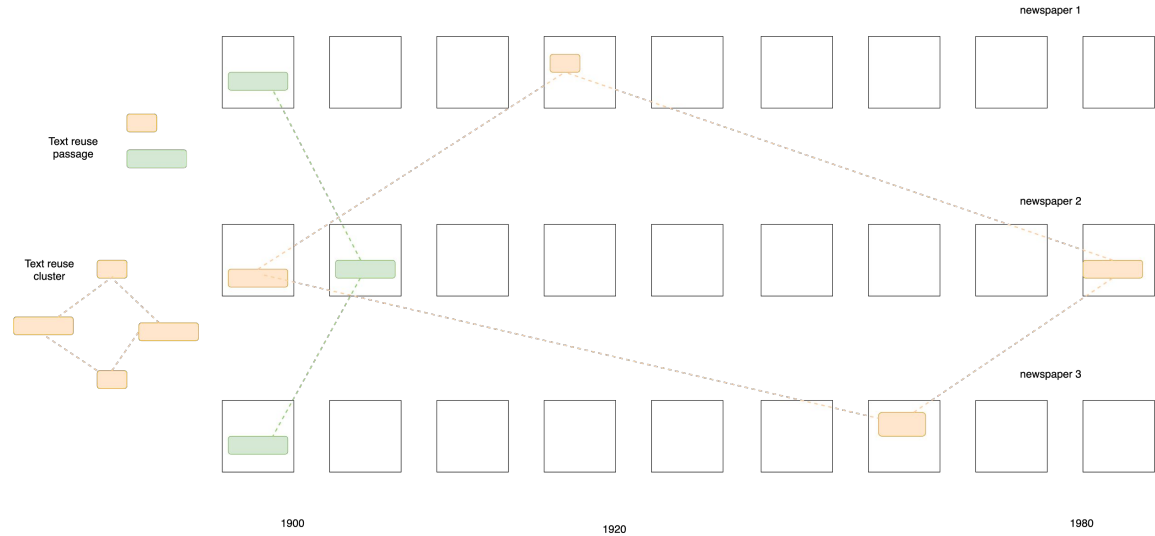
TEXT REUSE IN IMPRESSO

Detection, post-processing, exploration.

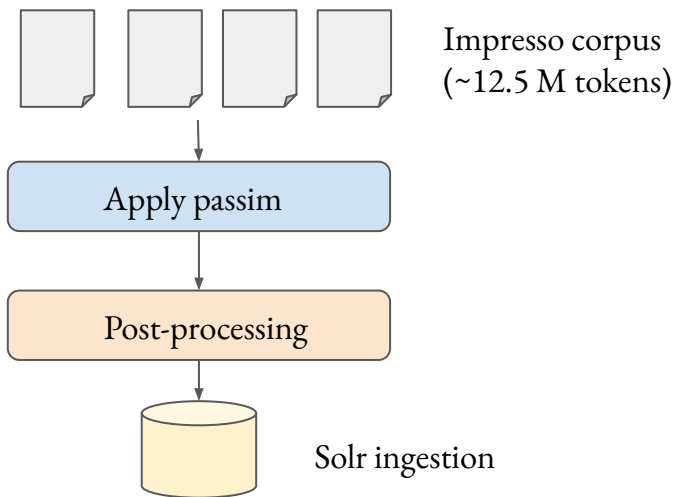
TR-RELATED JARGON

TR *cluster*: a group of *passages*, from different newspapers content items (~articles), that have a certain amount of text in common.

Several clusters can be detected within a single content item – esp. if a long one.
The length of TR passages can vary considerably.



TR IN IMPRESSO: WORKFLOW & SOME STATS



Post-processing of TR clusters:

- Cluster size
- Lexical overlap
- Time span (# of days)
- Connected clusters

Detecting Text Reuse with Passim
Matteo Romanello and Simon Hengchen

In this lesson you will learn about text reuse detection – the automatic identification of reused passages in texts – and why you might want to use it in your research. Through a detailed installation guide and two case studies, this lesson will teach you the ropes of Passim, an open source and scalable tool for text reuse detection.

Peer-reviewed CC-BY 4.0 Support PH

<https://doi.org/10.46430/phen0092>

6,177,815 clusters

16,099,821 passages

~17% of content items contain TR

TR IN THE IMPRESSO APP

Display of text reuse in the **reading mode**.

The screenshot displays the Impresso app interface for a newspaper page from Tuesday, March 14, 1911. The main content area shows the article "L'art de la réclame" with highlighted text reuse. The sidebar on the right provides a cluster summary and a timeline.

Cluster Summary:

- CLUSTER #c51539713128
- 59-57%
- 4 ARTICLES from TUE, MAR 14, 1911 to TUE, DEC 20, 1927
- LATEST SAMPLE: Bien qu'il ait mis, autant que quiconque, la main, à la plume, soutenu des polémiques retentissantes, publié les «Mémoires d'un Journaliste» qui sont un document de la vie parisienne d'il y a 40 et 50 ans, H. de Villemessant...
- Timeline: 1738 — 16 years — 2018

Text Reuse Analysis:

4 TEXT REUSE PASSAGES AVAILABLE

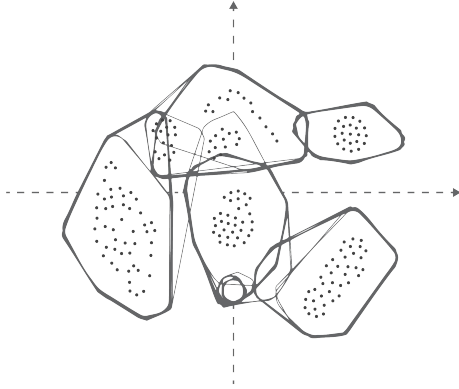
1 L'art de la réclame

2 Le grand art. — Un des maîtres du genre: H. de Villemessant. — Un mot de la fin. — «Au serin de Marie». — A New-York et à Chicago: la réclame vivante aux vitrines. — La dame de l'hôtel américain.

3 Bien qu'il ait mis, autant que quiconque, la main à la plume, soutenu des polémiques retentissantes, publié les «mémoires d'un journaliste» qui sont un document piquant de la vie parisienne

4

TR IN THE IMPRESSO APP



Text reuse explorer: filtering (even on a collection), overview of what's in a TR cluster, and which clusters are connected.

The screenshot shows the Impresso app interface. At the top, there's a search bar and navigation options like 'Newspapers', 'Topics', 'Inspect & Compare', and 'Text reuse'. The main content area displays a 'Cluster #c188978588270' with 2 articles from November 27, 1993. The cluster is defined by a lexical overlap of 94.57% and a cluster size between 1 and 54. Below the cluster information, there's a list of articles with their titles and dates. The first article is 'À LA VOLEE HOCKEY SUR GLACE Après les Russes Viatcheslav Bykov et Andrei Kholmutov (Fribourg-Gottéron), le HC Davos sera également renforcé par les Suédois Per Djos et Jan Larsson (Lugano) lors de la 67e Coupe Spengler, du 26 au 31 décembre. HOCKEY SUR GLACE Reto Pavoni (HC Klotten) était toujours leader du trophée Jacques Plante des meilleurs gardiens au terme de la dix-huitième journée. Tenant du titre, Pavoni précède d'un point son coéquipier en équipe nationale, Renato Tosio (HC Berne). Trophée Jacques-Plante. Classement après 18 journées : Reto Pavoni (Klotten) 23 points. 2. Renato Tosio (Berne) 22.3. Dino Stecher (Fribourg-Gottéron) 18.4. Lars Weibel (Lugano) 17.5. Olivier Anken (Bienne) 4.6. Patrick Schöpf (Zoug) 3. VOLLEYBALL Le circuit de la Coupe mondiale 1994, organisé du 6 mai au 12 juin pour la phase préliminaire, les 27 et 28 juillet pour la phase finale, sera doté de six millions de dollars que se partageront douze pays, répartis en trois groupes. La phase préliminaire sera organisée sur six weekends, chaque équipe rencontrant les trois autres en matches aller et retour. Les deux premiers de chaque groupe seront qualifiés pour le tournoi final organisé en Italie. Onze des douze participants figuraient déjà dans la Ligue mondiale 1993, le seul nouveau venu étant la Bulgarie qui a remplacé le Canada. La répartition des prix a été ainsi établie : pour la phase préliminaire, les équipes recevront respectivement dans l'ordre du classement, 215 000, 200 000, 200 000 et 200 000 dollars. Pour la phase finale, toujours dans l'ordre du classement : 1 000 000, 500 000, 300 000 et 200 000 dollars. Le meilleur joueur du tournoi recevra quant à lui 1 200 dollars. BOXE Le Valaisan de Genève Jean Chiarelli (28 ans), dans la catégorie des poids welter, et le Thurgovien Stefan Angehrn (29 ans), dans la catégorie des mi-lourds, participeront au traditionnel meeting de la Saint-Etienne, le 26 décembre, au Kursaal de Berne.

Handwritten notes and annotations on the left side of the page. The notes are organized into sections, some with red circles and arrows pointing to specific elements in the screenshot. The sections include:

- 3m. REUSE CLUSTERS**: A section with a red circle and an arrow pointing to the cluster information in the screenshot.
- CLUSTERS**: A section with a red circle and an arrow pointing to the cluster information in the screenshot.
- 14 ARTICLES**: A section with a red circle and an arrow pointing to the list of articles in the screenshot.
- 17 Felix Bonjour**: A section with a red circle and an arrow pointing to the article title in the screenshot.
- ES 75 Ann de Felix BONJOUR**: A section with a red circle and an arrow pointing to the article title in the screenshot.
- BOUZE FELIX BONJOUR**: A section with a red circle and an arrow pointing to the article title in the screenshot.

CLUSTER EXAMPLES

LA JOURNALISTE Lundi 4 août 1980 2

OFFRES D'EMPLOIS **OFFRES D'EMPLOIS**

Nous recherchons
de suite ou date à convenir

1 VENDEUSE
rayon ménage-mobilier

1 VENDEUSE
rayon charcuterie-fromages

1 VENDEUSE
auxiliaire (quelques heures par semaine)

Prière d'adresser vos offres aux
Grands magasins
Innovation SA, Payerne
ou prendre rendez-vous au 037/613333

GRANDS MAGASINS
Innovation
PAYERNE SA

J'engage
UN BOULANGER-PÂTISSIER QUALIFIÉ

Laboratoire moderne.
Avantages sociaux.
Ecrire de suite ou pour date à convenir.
Gaston REPOUD
Boulangier-pâtisseries
1638 La Tour-de-Tréme
☎ 029/272 88

ideal job

Nous sommes recrutés par une entreprise dynamique pour le recrutement d'un

SECRETARE DE DIRECTION
25-30 ans, de langue maternelle française, diplômée de l'enseignement secondaire et d'anglais, bonne culture générale, méticuleuse, travailleuse indépendante.

Vos tâches principales seront les suivantes: accueil du client, information des clients, gestion des stocks, etc.

Nous offrons une collaboration active, un travail intéressant ainsi que des possibilités d'évolution.

Adresser directement vos candidatures à M. Baudouin Claude
☎ 029/11313

JEUNE SCOLIERE
pour accompagnement scolaire, surveillance de la classe.

Adresser vos lettres à M. Baudouin Claude
☎ 029/11313

ideal job

Vous êtes le principal responsable de la fabrication des produits.

Imprimerie Saint-Paul

EMPLOYÉ DE GARAGE
pour le service service et pour la préparation des véhicules.

Garage Central Quincaillerie S.A.
Avenue NORD, 10100
10100

M
NECESSAIRE
FRANCAIS

CUISINIER
pour le Restaurant de haut niveau AVEY

Nous offrons:
- salaire indexé de 45 heures;
- possibilité de travailler en équipe.

LA VILLE DE FRIBOURG
merit au concours le poste de
Préposé à l'aumônerie

Conditions:

- Être apte à maîtriser les problèmes de l'introduction, du fonctionnement et de la coordination de l'enseignement que dans une administration communale.

- Affiliation obligatoire à la caisse de prévoyance du personnel de l'Administration communale.

- Ecrire en fonction: le plus tôt possible.

Les offres de service accompagnées du curriculum vitae, photographie, photocopies de certificats et diplômes, doivent être adressées jusqu'au 23 août 1980, au Secrétaire de Ville, Maison de Ville, 1700 Fribourg, où le cadre des charges peut être consulté.

économiser
est la publicité
c'est vouloir
recueillir
sans avoir
rien

Partir bien reposé
tcs) Pas à la sortie du travail ou après trois à quatre heures de sommeil seulement.

INGENIEUR ETS
pour la construction de bâtiments industriels et commerciaux.

LA VILLE DE FRIBOURG
merit au concours le poste de
Préposé à l'aumônerie

Conditions:

- Être apte à maîtriser les problèmes de l'introduction, du fonctionnement et de la coordination de l'enseignement que dans une administration communale.

- Affiliation obligatoire à la caisse de prévoyance du personnel de l'Administration communale.

- Ecrire en fonction: le plus tôt possible.

Les offres de service accompagnées du curriculum vitae, photographie, photocopies de certificats et diplômes, doivent être adressées jusqu'au 23 août 1980, au Secrétaire de Ville, Maison de Ville, 1700 Fribourg, où le cadre des charges peut être consulté.

Imprimerie Saint-Paul



Reprinted 68 times over 21 years (1977-1998), in 3 Swiss newspapers.

Content item type:

- Advertisement (n = 64)
- Article (n = 4)

Cluster [133801](#)

(Filter by type not supported by the impresso app.)

CLUSTER EXAMPLES

8 TR passages, with 7.5% lexical overlap. 100% advertisements.

Job postings by a consultancy company. Texts differ, but similar *template*.

The screenshot displays a web interface for text reuse clusters. At the top, it shows 'Cluster #c21855' with 8 articles from TUE, MAR 4, 1997 to TUE, NOV 11, 1997 (1 year) with 7.5% lexical overlap. Below this are navigation options: 'OVERVIEW', '8 PASSAGES', and '3 CONNECTED CLUSTERS'. A dropdown menu is set to 'DEFAULT'. Three job postings are listed, each with a small icon and a 'NaN' similarity score. The first posting is for a 'COMPTABLE' position at 'Agence Raymond Pont' in Geneva, dated Tuesday, June 24, 1997. The second is for a 'SECRETAIRE-ASSISTANTE DE GESTION français-anglais' at 'RAYMONDE PONT CONSEILS D'ENTREPRISES ET SELECTION DE PERSONNEL' in Geneva, dated Tuesday, November 11, 1997. The third is for a 'SECRETAIRE TRILINGUE français-anglais-allemand' at 'RAYMONDE PONT CONSEILS D'ENTREPRISES ET SELECTION DE PERSONNEL' in Geneva, dated Tuesday, March 4, 1997. All three postings follow a similar template structure, including company name, address, and contact information.

TEXT REUSE BEYOND IMPRESSO

Inspirations from past and current projects.

GRAPH – TEXT REUSE IN RARE BOOKS

ETH [project](#), interface development by Benoit Seguin.

Text corpus: OCR of 1,300 rare books on architecture from ETH library.

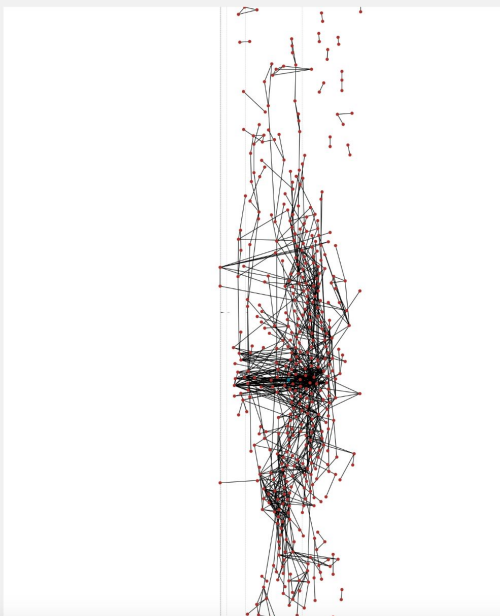
TRD performed with *passim*.

3 types of TR visualizations:

1. Graph
2. Matches
3. Compare
4. Flows

Interactive graph visualization of reuse between books, rearrangeable according to time (x-axis).


ETH zürich | Graph – Text reuse in rare books




Filter graph by title or author (563)

Order graph with time Hide ulterior editions of the same book (experimental)

[The grecian orders of architecture](#)

	Year	1768
	Authors / Collaborators	Riou, Stephen
	Imprint	London : printed by J. Dixwell, 1768
	Localisation	ETH Library, Rar 1408 GF
	Description	[10] Bl., 78 S. : Ill. (18 Kupfertafeln) ; 46 cm
	Language	English
	Nb pages	151


SEE ON  [raa](#)



GRAPH – TEXT REUSE IN RARE BOOKS

TR clusters are grouped by individual publications.

≡ **ETH** zürich | Graph – Text reuse in rare books


Nuova raccolta d'autori che trattano del moto dell'acque



	Year	1766
Authors / Collaborators		
Imprint	Parma : per Filippo Carmignani, 1766-1786	
Localisation	ETH Library, Rar 4652	
Description	7 volumi ; 22 cm	
Language	Italian	
Nb pages	3542	

SEE ON  [rara](#) 

Detected text reuses with 6 books

Architettura d'acque

	Year	1699
Authors / Collaborators	Barattieri, Giovan Battista [1600-1677]	
Imprint	Piacenza : nella Stampa Ducale di Lealdo Leandro Bazachi, 1699	
Localisation	ETH Library, Rar 687	
Description	2 parti ; 33 cm	
Language	Italian	
Nb pages	618	

SEE ON  [rara](#) 

3 types of TR visualizations:

1. Graph
2. Matches
3. Compare
4. Flows

[\(Link to match\)](#)

GRAPH – TEXT REUSE IN RARE BOOKS

TR is highlighted directly on the page facsimile.

3 types of TR visualizations:

1. Graph
2. Matches
3. Compare
4. Flows

The screenshot displays a digital library interface for the ETH Zürich collection. The main content area shows two facsimile pages from a book titled "DELLA MISURA DELL'ACQUE CORRENTI DI D. BENEDETTO CASTELLI MONACO CASSINENSE." The left page features a large block of text with a red highlight. The right page shows a similar page with an illustration of a landscape and a smaller red highlight. The interface includes a search bar at the top, navigation buttons, and a list of thumbnails on the sides.

GRAPH – TEXT REUSE IN RARE BOOKS

(Static) visualisation of TR flows across different sections of any two matching books.

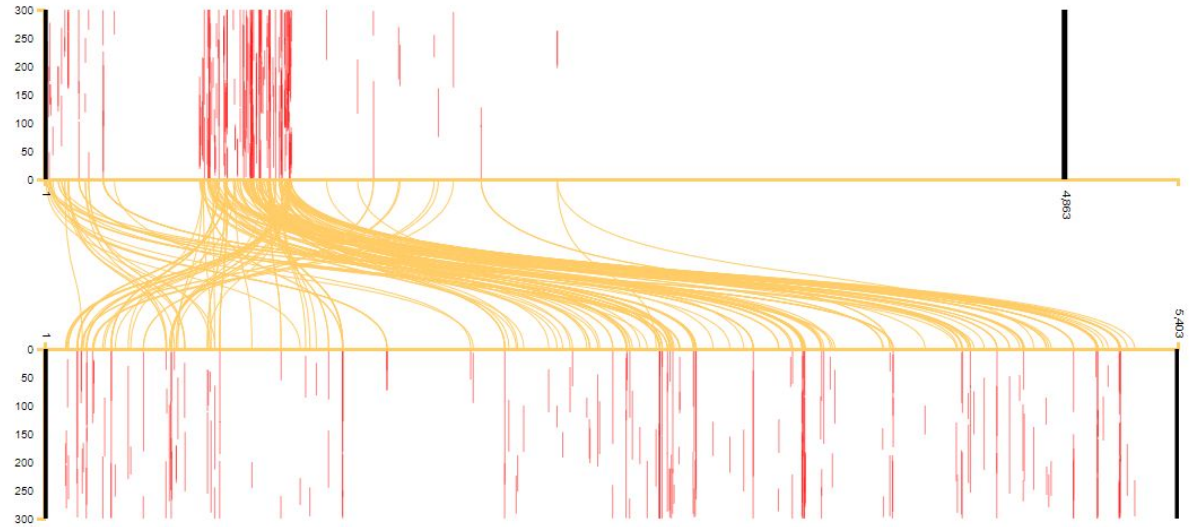
3 types of TR visualizations:

1. Graph
2. Matches
3. Compare
4. **Flows**



KITAB

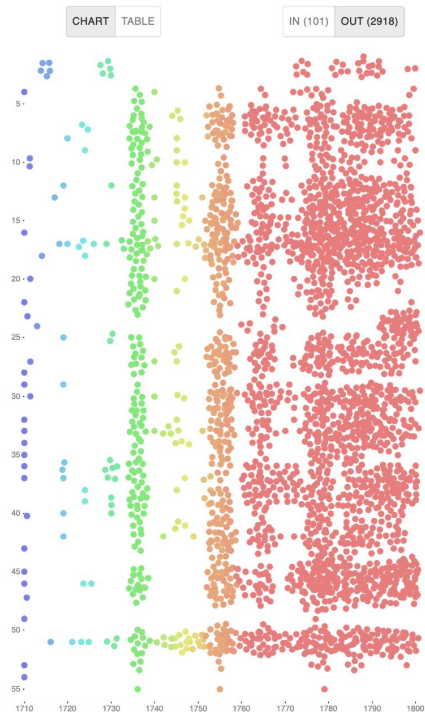
[Kitab](#) project.



RECEPTION READER

App to explore text reuse in the ECCO corpus.

Developed at the Helsinki Computational History Group ([COMHIS](#)).



Shakespeare, William (1564-1616)
1710 - Macbeth

M A C B E T H

Ban. The Earth has bubbles like the Water,
And these are fons of them: How soon they are vanish'd!
Macb. Th' are run'd to Ainy what form'd Corporal
Is melted into nothing, would they had fluid.
Ban. Were such things here as we discours'd of now?
Or have we talk'd some infectious Herb
That exasperates our Rage?
Macb. Your Children shall be Kings.
Ban. You shall be King.
Macb. And Thine of Cawdor too, were it not for
Ban. Just to that very tune! who's here?
Enter Macduff.
Macb. Macduff, the King has happily receiv'd
The news of your success: And when he reads
Your personal Venture in the Rubick filds,
His Wonder and his Praise then comes
Which shall exceed when he reviews your worth.
Not faring as the Images of Death
Made by your fell's each Meffenger which came
Being laden with the Prizes of your Valour,
Some'd post to feak your Glories to the King;
Who, for an earnest of a greater Honour,
Hath sent from him, to call you Thine of Cawdor:
In which Addition, High, most Noble Thane!
Ban. What, can the Devil speak true?
Macb. The Thane of Cawdor lives;
Why do you dress me in his borrow'd Robes?
Macb. 'Tis true, Sir; He, who was the Thane, lives yet;
But under heavy Judgment bears that life
Which he in justice condemn'd to lose.
Whether he was combin'd with those of Norway,
Or did lift the Rebel's standard,
Or whether he concurr'd with both, to cause
His Country's danger, Sir, I cannot tell:
But, Treason Capital, confid and prov'd,
Have overthrow him.
Macb. Glens and Thane of Cawdor!
The greatest is behind. My noble Partner!
Do you not hope your Children shall be Kings?
When those who gave to me the Thane of Cawdor
Promis'd to let to them.
Ban. If all be true,
You have a Title to his Crown, as well
As to the Thane of Cawdor. It from flatterer:
But many times, to win us to our harm,
The Instruments of darkness tell us Truths,
And tempt us with false smiles, that they may

Page 8

Griffin, Benjamin (1680-1740)
1720 - Whig and Tory

The Dedication.

tion: The great Things You
have said and done, in our
Senate and our Camps, in the
great Cause of Liberty, are
lasting Trophies to eternize
your Fame: Our Annals are
full of 'em;

And when we read
Your Personal Venture in Your Country's Cause,
Our Wonder and our Praises then contend
Which shall exceed: When we review Your Worth,
We find You in the stout Britannick Ranks,
Not starting at the Images of Death,
Made by Your self. Shake.

'T W E R E endless, should I
attempt to recount Your noble
ACTS; in whose Praise I
might be lavish, did I not
fear to offend Your Lord-
ship's Modesty; to whom
this *Motto* may with the great-
est Justice be applied:
Non magna loquimur, sed vivimus.
It being Your Lordship's pro-
per

Page 5

[Reception reader](#) app: reuses of Shakespeare's *Macbeth*

SCHOLARLY RECEPTION IN CLASSICS

Quotations (and references) extracted from JSTOR, and used as a proxy for scholarly reception.

Cited Loci :: Aeneid Home Explore About

Display: quotations

In Focus: Book 1, lines 1-50

Results: quotations

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1-50	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
51-100	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
101-150	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
151-200	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
201-250	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
251-300	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
301-350	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
351-400	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
401-450	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
451-500	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
501-550	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
551-600	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
601-650	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
651-700	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
701-750	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
751-800	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
801-850	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
851-900	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
901-950	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
951-1000	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue

Arma virumque cano, Troiae qui primus ab oris
Italiam, fato profugus, Laviniaque venit
litora, multum ille et terris iactatus et alto
vi superum saevae memorem Lunois ob iram;
5 multa quoque et bello passus, dum conderet urbem,
inferretque deos Latio, genus unde Latinum,
Albanique patres, atque altae moenia Romae.
Musa, mihi causas memora, quo numine laeso,
quidve dolens, regina deum tot volvere casus
10 insignem pietate virum, tot adire labores
impulerit. Tantaene animis caelestibus irae?
Urbs antiqua fuit, Tyrii tenuere coloni,
Karthago, Italiam contra Tiberinaque longe
ostia, dives opum studiisque asperrima belli;
15 quam Iuno fertur terris magis omnibus unam
posthabita coluisse Samo; hic illius arma,
hic currus fuit; hoc regnum dea gentibus esse,
si qua fata sinant, iam tum tenditque fovetque.

Aen. 1.1 112

"We See by the Papers"
RICHARD M. FRAZER, <SUFFIX>JR.</SUFFIX>
The Classical Journal 1963
DOI: 10.2307/3294648

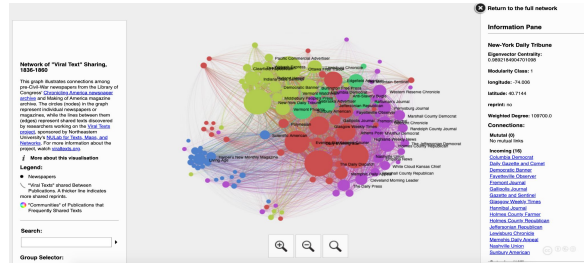
occasionally to introduce a scientific paper. In one such application, Dr. Kenneth S. Cole of the National Institute of Neurological Diseases and Blindness, chose the opening lines of Vergil's *Aeneid*, "*Arma Virumque Cano*," t introduce a paper presented at the first International Biophysics Congress, held last year. When the paper, "The Advance of Electrical Models for Cells and Axons," was published in the *Biophysical Journal*,

— (quotation:
ad29e2286d7b78396ae8291b219c7c1a)

Cited *loci* of the *Aeneid*: [interactive visualisation](#).

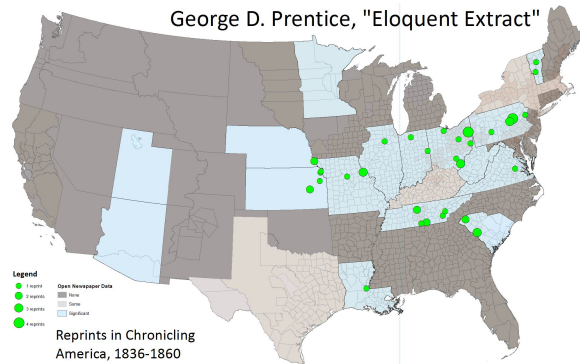
VIRAL TEXTS

Corpus of pre-1861 American newspapers in Library of Congress' *Chronicling America* archive.



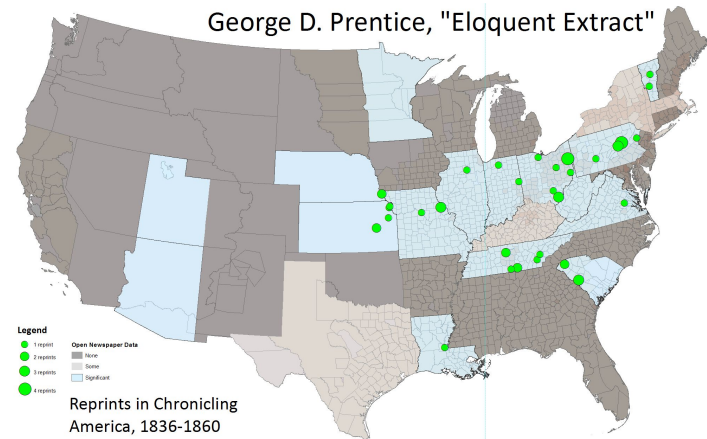
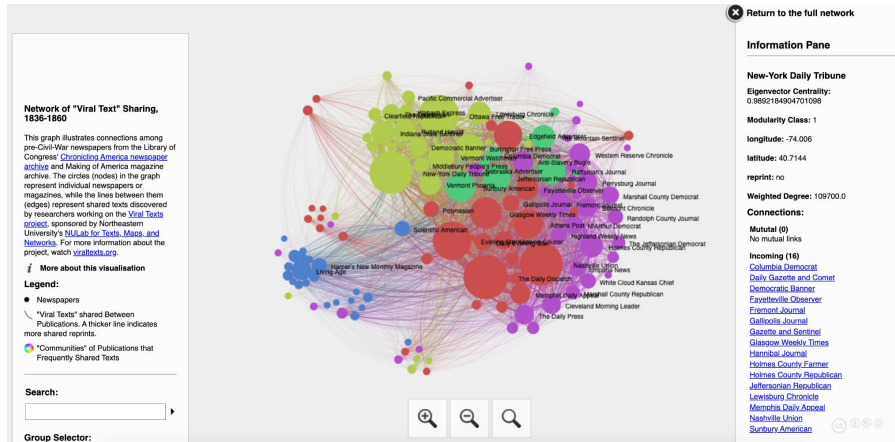
Historical research insights (Cordell 2015):

- The **network author model** allowed for a better understanding of 19C newspaper literature → something “composed incrementally, by a community of writers and editors in a network.”
- Network analysis pointed to newspapers in understudied cities as “important brokers of textual exchange”.



VIRAL TEXTS

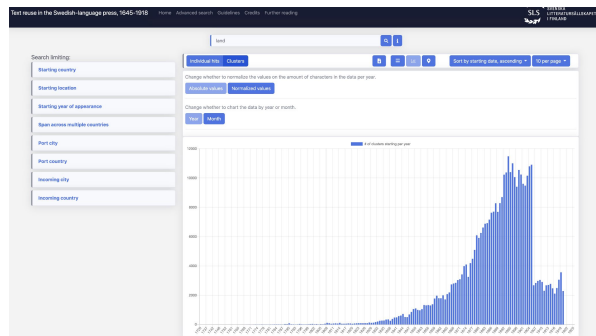
Corpus of pre-1861 American newspapers in Library of Congress' *Chronicling America* archive.



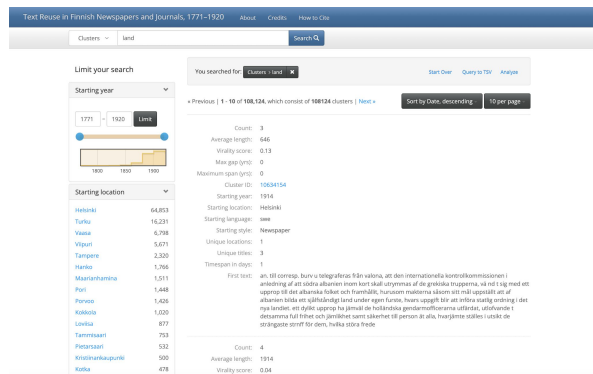
(from Cordell 2015)

<http://networks.viraltexts.org/1836to1860/index.html>

SCANDINAVIAN TEXT REUSE



[Text reuse in the Swedish-language press, 1645-1918](#)



[Text Reuse in Finnish Newspapers and Journals, 1771-1920](#)

TRD performed with BLAST (Vesanto et al. 2017)

For each TR cluster, a *virality score* is computed, which “is a way of approaching how efficiently a particular text is circulated through the media network” ([source](#)).

Virality score = $\text{number_venues} * \text{number_countries} * (1/\text{time_span})$.
The score is then normalized to a 0-100 range.

TR clusters are characterised with further properties:

- Max. gap in the circulation chain (# of days)
- Start location / port location / incoming location
- Crossed (boolean)
- Average length

TEXT REUSE AT SCALE

Some ideas for this workshop.

NEEDLE-IN-THE-HAYSTACK PROBLEM



S. Sachsalber, *Inside*, Paris 2014. (image [source](#))

Impresso TR data: very large and still too noisy to be analysed.

Double strategy:

- Provide effective ways to filter out uninteresting things
- Make it easier for the user to spot the interesting things (e.g. long term reuse, viral articles in a collection)

Criteria for filtering:

- Average passage length (of a cluster)
- Lexical overlap (find a meaningful threshold)
- Predominant type of content items
- ...

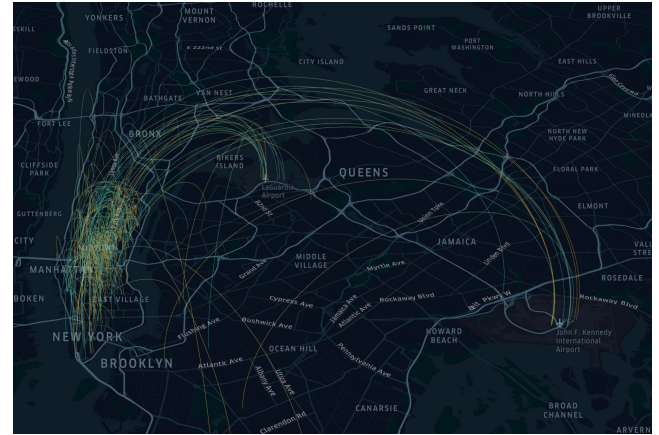
VIRALITY IN IMPRESSO

- Virality: unharvested low-hanging fruit
- Suitability of impresso corpus (?)
- Tension: generic vs specialised interface

Historical research questions:

- What type of texts circulated virally *across* countries as opposed to *within* countries?
- Which newspapers/cities/cantons played a primary role in spreading viral news?
- Similarities/differences wrt (re)printing cultures in other countries (FI, US)?

Visualisation 💡 Geographical map, with arcs for flows of news (reprints) + time-based animation.



Arc layer in Kepler.gl

([source](#))

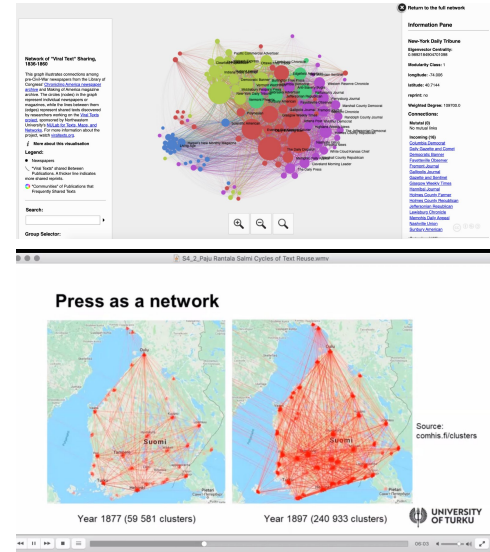
NETWORKS OF TEXT REUSE

Currently, **no high-level “picture”** of reuse dynamics in our corpus.

Graph-based visualisations for impresso’s TR data could help:

- *Cluster level* → clarify genealogical relationships between clusters
- *Corpus level* → explore the network of reuse (hubs, communities)

Characterisation of reuse relationships through semantic enrichments (topics)



REFERENCES

- Smith, D.A., Cordell, R., Mullen, A., 2015.** Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History* 27, E1–E15. <https://doi.org/10.1093/alh/ajv029>
- Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., Ginter, F., 2017.** Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910, in: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. pp. 54–58.
- Cordell, R., 2015.** Reprinting, Circulation, and the Network Author in Antebellum Newspapers. *American Literary History* 27, 417–445. <https://doi.org/10.1093/alh/ajv028>
- Salmi, H., Paju, P., Rantala, H., Nivala, A., Vesanto, A., Ginter, F., 2020.** The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 0, 1–15. <https://doi.org/10.1080/01615440.2020.1803166>
- Büchler, M., Burns, P.R., Müller, M., Franzini, E., Franzini, G., 2014.** Towards a Historical Text Re-use Detection, in: *Biemann, C., Mehler, A. (Eds.), Text Mining, Theory and Applications of Natural Language Processing*. Springer International Publishing, pp. 221–238.

REFERENCES

Bär, D., Zesch, T., Gurevych, I., 2012. Text Reuse Detection using a Composition of Text Similarity Measures, in: Proceedings of COLING 2012. Presented at the COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India, pp. 167–184.

Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C.W., Ossewaarde, R., Jacobson, S.L., 2012. The Tesseract Project: intertextual analysis of Latin poetry. Lit Linguist Computing. <https://doi.org/10.1093/lit/fqs033>

Moritz, M., Steding, D., 2018. Lexical and Semantic Features for Cross-lingual Text Reuse Classification: an Experiment in English and Latin Paraphrases, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Presented at the LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan.

Scheirer, W., Forstall, C., Coffee, N., 2016. The sense of a connection: Automatic tracing of intertextuality by meaning. Digital Scholarship in the Humanities 31, 204–217. <https://doi.org/10.1093/dsh/31.2.204>

Romanello, M., Hengchen, S., 2020. Detecting Text Reuse with Passim. The Programming Historian. <https://doi.org/10.46430/phen0092>