Passim vs Tracer

	Tracer	**Passim**
Language	Java	Java / Scala
Compiler	ant	sbt
Dependencies	• java8 • medusa (comes together with tracer)	• java8 • apache spark • hadoop
Cross-platform?	Yes	No
Restrictions	—	Does't work on OS X (due to Hadoop's nativelib)
Open-source?	Partly; the repo is restricted to registered users, but they welcome contributions	Yes
Source code	eTRAP's GitLab http://vcs.etrap.eu/tracer-framework/tracer	GitHub https://github.com/dasmiq/passim
Manual	https://tracer.gitbook.io/-manual/ Good detailed manual, but still incomplete. Also, after compiling Tracer you can generate a javadoc with a short description of every class used in the program.	The only manual is a README in Passim's GitHub repo, but it's not too bad. No minimal working example.
Community	eTRAP, active community	Used in several projects like Kitab, but no active community
Support	You have to register on http://www.etrap.eu/redmine/login and wait for account approval to raise an issue. Approval seems to never happen (still pending a month after account creation)	Very easy to open an issue on GitHub, but they aren't answered.
Built-in visualisation	Yes (TraVIZ)	No

Built-in alignment	Yes Outputs aligned passage ids together with their absolute overlap (nr of shared features) to the *.link* file	Yes Produces results in *.json* or *.parquet* format, which contain a document/line from the right-hand, "target" text, along with information about corresponding passages in the left-hand, "witness" or "source" text.
Algorithm/scoring	Their own similarity score	Single-link(age) clustering

Passim

- D. A. Smith, David R. Cordell, and David A. Ryan Abby Mullen. 2015. Computational methods for uncovering reprinted texts in antebellum newspapers. American Literary History, 27: E1 – E15 (pdf)
- MacLaughlin, Ansel, and David A. Smith. "Content-based Models of Quotation." (pdf)
- Romanello, Matteo. *Detecting Text Reuse in Newspapers Data with Passim*. No. POST_TALK. 2018. (slides)
- https://github.com/dasmiq/passim
- Eclipse Public License

Requirements & installation

1. Java 8 (not higher!)
2. Scala
3. Sbt (a build tool for JVM languages)
    a. **NB!** Build Passim with sbt before installing Apache-Spark. Spark comes with Java 11, which results in build conflicts when you are compiling Passim with sbt.
4. Apache-Spark
    a. Comes with Java 11
    b. Depends on Hadoop, which has native libraries not supported in OS X →
        i. Won’t run on a Mac

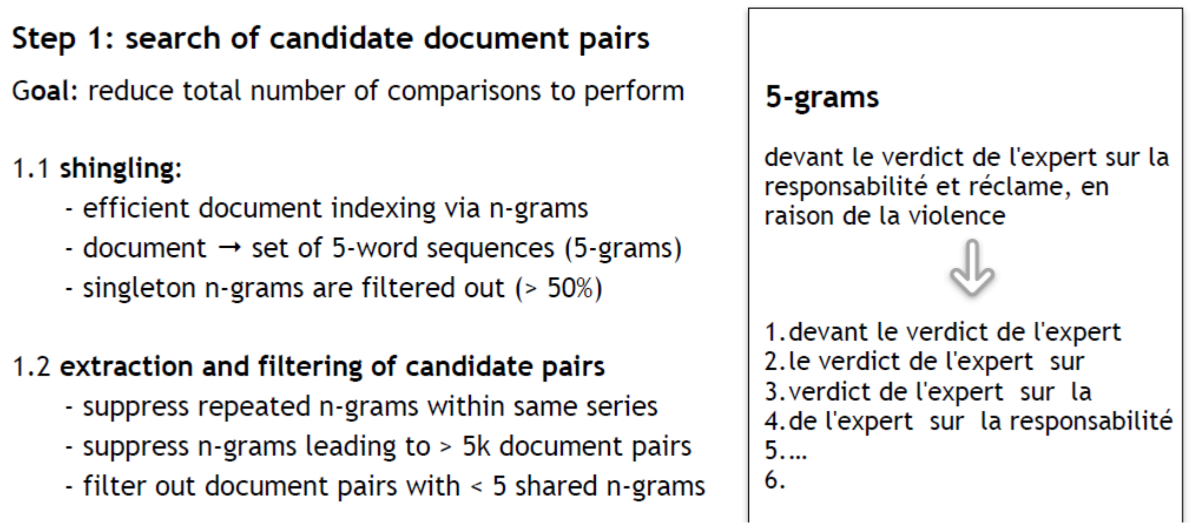
Key features (from documentation)

- Detects and aligns similar passages in text.
- Can be run either in query mode, to find quoted passages from a reference text, or all-pairs mode, to find all pairs of passages within longer documents with substantial alignments.
- The input to passim is a set of *documents* (whole books, pages of books, whole issues of newspapers, individual newspaper articles, etc.) Minimally, a document consists of an identifier string and a single string of text content.
- Sometimes it is useful to group documents into *series*. Text reuse within the series will be ignored.

- Output: clusters and pairwise alignments between all matching passages
- Uses the Smith–Waterman alignment algorithm to find all pairs of passages within longer documents (source and derived)
- Parameters:

Parameter	Default value	Description
`--n`	5	N-gram order for text-reuse detection
`--max-series`	100	Maximum document frequency of n-grams used.
`--min-match`	5	Minimum number of matching n-grams between two documents.
`--relative-overlap`	0.5	Proportion that two different aligned passages from the same document must overlap to be clustered together, as measured on the longer passage.

Algorithm



Step 1: search of candidate document pairs

Goal: reduce total number of comparisons to perform

1.1 **shingling:**
- efficient document indexing via n-grams
- document → set of 5-word sequences (5-grams)
- singleton n-grams are filtered out (> 50%)

1.2 **extraction and filtering of candidate pairs**
- suppress repeated n-grams within same series
- suppress n-grams leading to > 5k document pairs
- filter out document pairs with < 5 shared n-grams

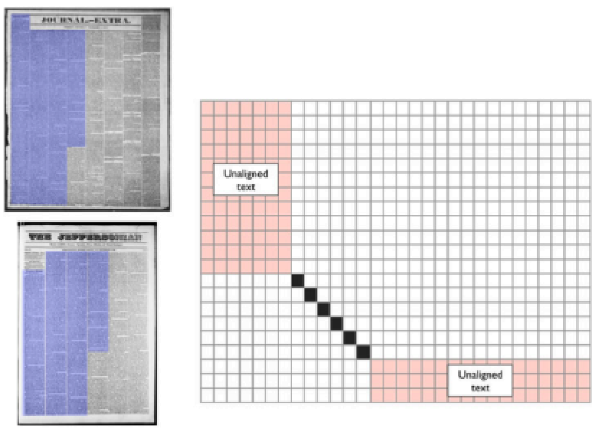
5-grams

devant le verdict de l'expert sur la responsabilité et réclame, en raison de la violence

1. devant le verdict de l'expert
2. le verdict de l'expert sur
3. verdict de l'expert sur la
4. de l'expert sur la responsabilité
5. ...
6.

Step 2: local document alignment

Goal: given a set of document pairs, obtain a set of aligned passage pairs



*source of illustration: Smith et al. 2015, Fig. 4, p. E9

Step 3: passage clustering

Goal: group similar passages into clusters

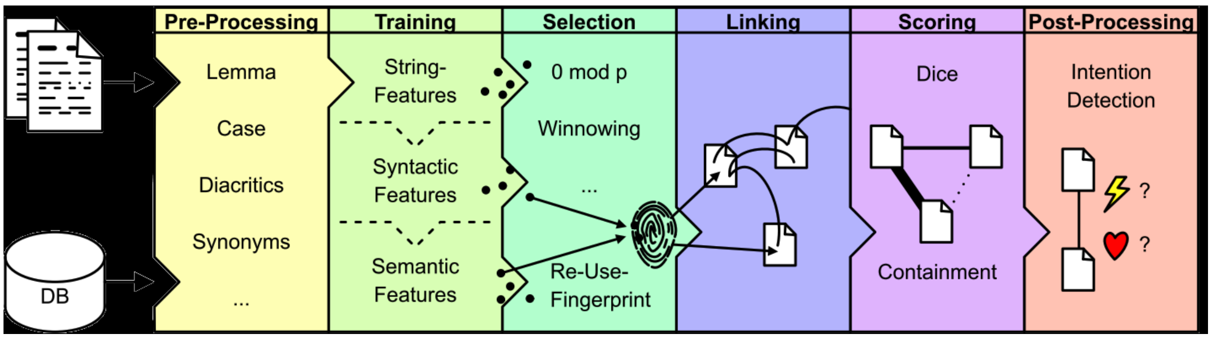
- "single-link" clustering
- overlap threshold 50%

Single-link(age) clustering

- A method of hierarchical clustering. It is based on grouping clusters in bottom-up fashion (agglomerative clustering), at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other (wiki)
- The similarity of two clusters is the similarity of their *most similar* members (Stanford NLP)

Tracer

- **Büchler, M.**, Burns, P. R., Müller, M., **Franzini, E.**, **Franzini, G.** (2014) 'Towards a Historical Text Re-use Detection', In: Biemann, C. and Mehler, A. (eds.) *Text Mining, Theory and Applications of Natural Language Processing*. Springer International Publishing Switzerland. (pdf)
- Documentation: https://gfranzini.gitbooks.io/tracer/content/
- Academic Free License 3.0 (AFL)



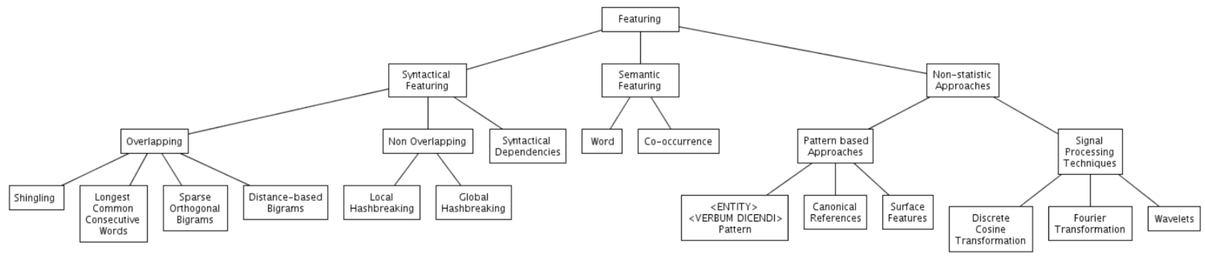
Requirements & installation

1. Any OS type (Windows/OS X/Linux)
2. Java 8 (not higher!)
3. Apache Ant (a build tool)
   a. Only use 'ant' command when compiling Tracer, using it with any flags (like "verbose") will later result in "Unable to access jarfile" error.
4. Troubleshooting: https://gfranzini.gitbooks.io/tracer/content/support/troubleshooting/

Key features (from documentation)

- A framework of ca. 700 algorithms, whose features can be combined to create the optimal model for detecting those words, sentences and ideas that have been reused across texts
- Helps to identify plagiarism in a text, as well as verbatim and near verbatim quotations, paraphrase and even allusions
- reuse results *can* be visualised in a more readable format via TRAViz

- **language independent** and has been successfully tested on Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin and Tibetan
- Supports intralinking (reused passages within a document) and interlinking (matches between different documents)
- N-gram features, can be characters or words, can be pruned and weighed
    - Overlapping (shingling)
    - Non-overlapping (hash-breaking)
    - Distance-based (good for detecting paraphrases)



Scoring

The *Scoring* step of TRACER assigns a weight to a reuse pair based on an internal scoring metric. The resemblance score is calculated as follows:

$$\theta^{R'}_{\Theta}(s_i, s_j) = \theta^{O}_{\Theta}(s_i, s_j) \cdot \theta^{R}_{\Theta}(s_i, s_j) = \frac{|\vec{s_i} \cap \vec{s_j}|}{|\vec{s_i} \cup \vec{s_j}|}$$

The resemblance score *θ* (theta) is the quotient of the reuse overlap (∩) and the reuse union (∪), where the reuse overlap represents the digital fingerprint of reuse units that appear in two candidate strings/lines, and the reuse union represents the number of all digital fingerprints.

Experiments

A GitHub repository with preprocessed data and results (both Cath Almaine and Welsh law): https://github.com/ancatmara/text-reuse-test

Could only test Tracer, because Passim doesn't compile on Mac, and trying to run it from a Linux virtual machine under Mac/Windows ended up in memory error. An actual Linux server/a powerful laptop with Linux as the main OS is required to test Passim.

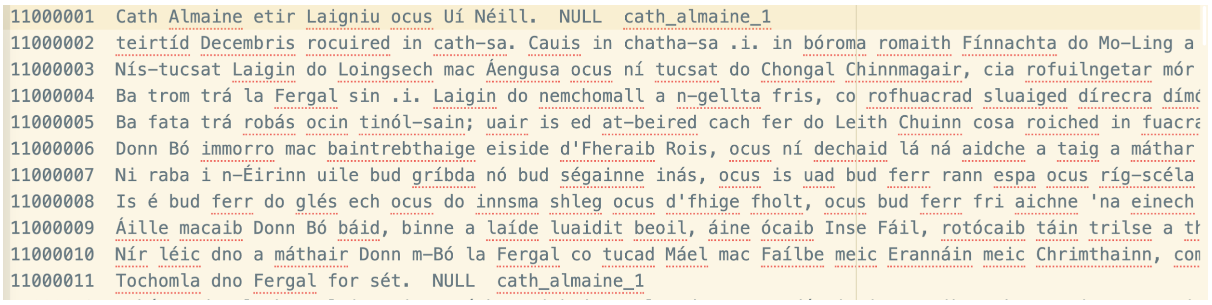
Tracer

```
java -Xmx600m -Deu.etrap.medusa.config.ClassConfig=conf/tracer_config.xml -jar tracer.jar > FILENAME.OUT
```

- Your data must be a single .txt file in tracer/data/corpora/*some-folder*
- Check and update tracer/conf/tracer_config.xml before running

Data format

1. The first column contains unique sentence IDs.
2. The second column contains the sentence.
3. The third column can contain the date of the file creation in the YYYY-MM-DD format or the word NULL. If you use NULL, make sure it is written in upper-case.
4. The fourth column contains the book or section from which the sentence is taken. This information is crucial for the visualisation shown in the Text Reuse Browser figure in postprocessing. The top drop-down menu you see there will list the information you provide in this fourth column.

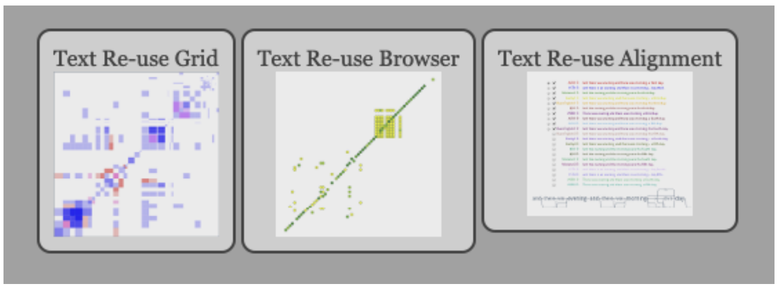


```
11000001  Cath Almaine etir Laigniu ocus Uí Néill.  NULL  cath_almaine_1
11000002  teirtíd Decembris rocuired in cath-sa. Cauis in chatha-sa .i. in bóroma romaith Fínnachta do Mo-Ling a
11000003  Nís-tucsat Laigin do Loingsech mac Áengusa ocus ní tucsat do Chongal Chinnmagair, cia rofuilngetar mór
11000004  Ba trom trá la Fergal sin .i. Laigin do nemchomall a n-gellta fris, co rofhuacrad sluaiged dírecra dímd
11000005  Ba fata trá robás ocin tinól-sain; uair is ed at-beired cach fer do Leith Chuinn cosa roiched in fuacra
11000006  Donn Bó immorro mac baintrebthaige eiside d'Fheraib Rois, ocus ní dechaid lá ná aidche a taig a máthar
11000007  Ni raba i n-Éirinn uile bud gríbda nó bud ségainne inás, ocus is uad bud ferr rann espa ocus ríg-scéla
11000008  Is é bud ferr do glés ech ocus do innsma shleg ocus d'fhige fholt, ocus bud ferr fri aichne 'na einech
11000009  Áille macaib Donn Bó báid, binne a laíde luaidit beoil, áine ócaib Inse Fáil, rotócaib táin trilse a th
11000010  Nír léic dno a máthair Donn m-Bó la Fergal co tucad Máel mac Faílbe meic Erannáin meic Chrimthainn, com
11000011  Tochomla dno Fergal for sét.  NULL  cath_almaine_1
```

TraViz

A built-in visualisation package for Tracer

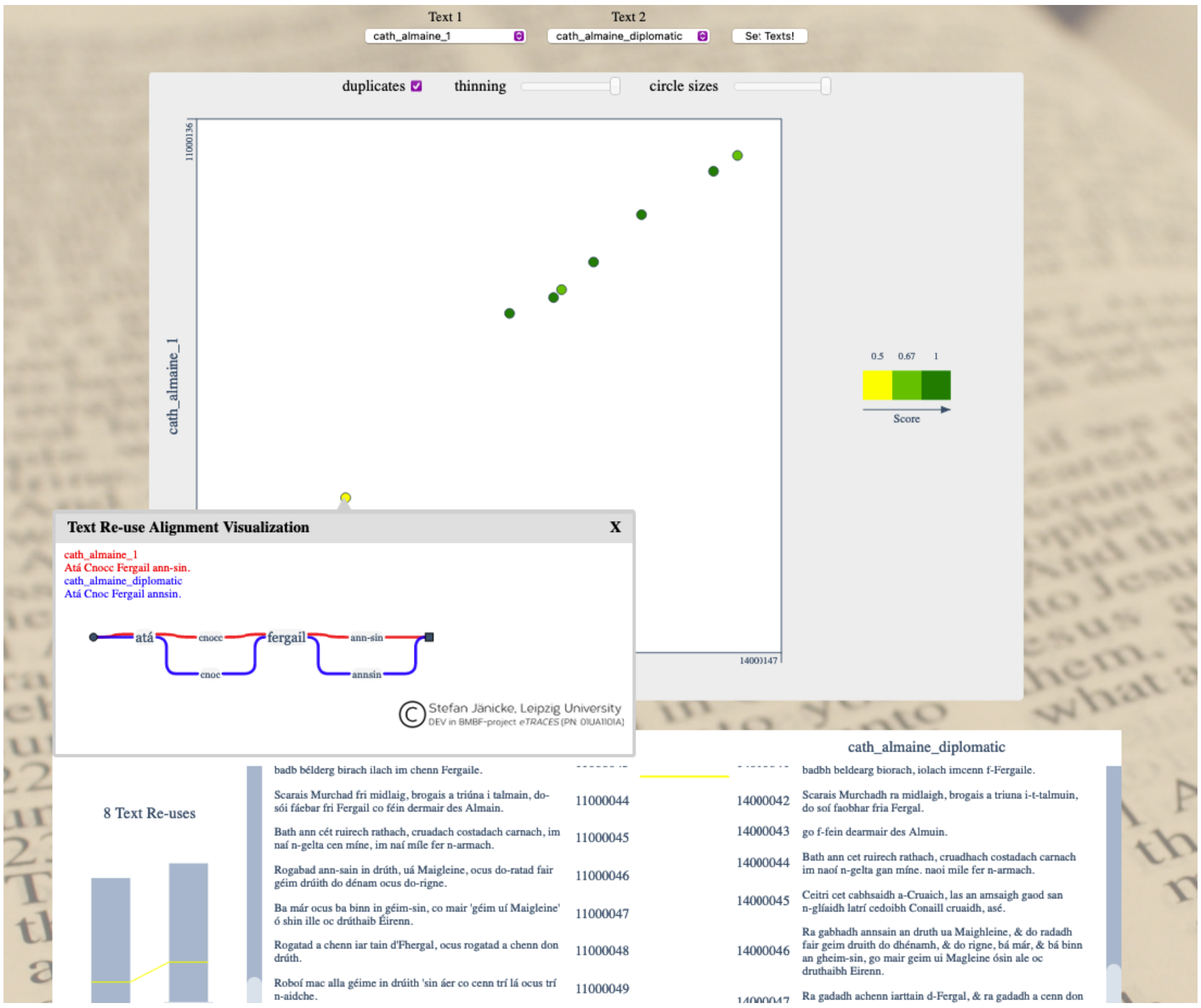
- Text re-use grid and alignment only work for comparing two texts; they seem not to work at all if there were more than 2 texts in the experiment
- Text re-use browser is also binary, but you can select the pair of texts to visualise



Cath Almaine Experiment

- 4 different editions, one of them diplomatic
- Minimal unit for comparison — a sentence (semi-automatic splitting)
- Basic TRACER configuration
  - No synonyms
  - No lemmatisation
  - No syntactic parsing
  - No diacritic removal

Text reuse browser in Traviz



Text 1: cath_almaine_1
Text 2: cath_almaine_diplomatic

duplicates ☑ thinning circle sizes

Text Re-use Alignment Visualization X

cath_almaine_1
Atá Cnocc Fergail ann-sin.
cath_almaine_diplomatic
Atá Cnoc Fergail annsin.

atá fergail

Stefan Jänicke, Leipzig University
DEV in BMBF-project *eTRACES* (PN: 01UA1101A)

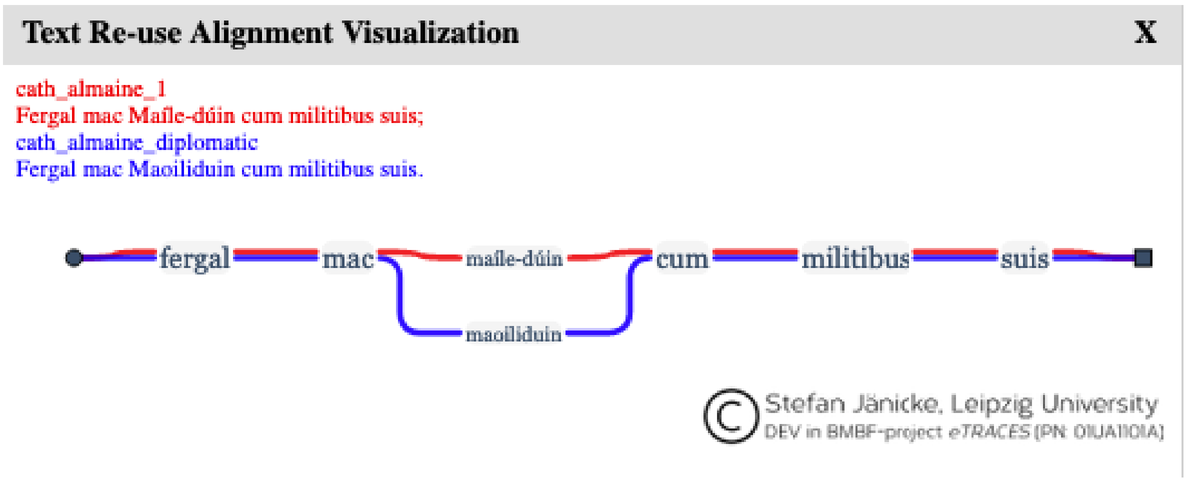
8 Text Re-uses

badb bélderg birach ilach im chenn Fergaile.	
Scarais Murchad fri midlaig, brogais a triúna i talmain, do-sói fáebar fri Fergail co féin dermair des Almain.	11000044
Bath ann cét ruirech rathach, cruadach costadach carnach, im naí n-gelta cen míne, im naí míle fer n-armach.	11000045
Rogabad ann-sain in drúth, uá Maigleine, ocus do-ratad fair géim drúith do dénam ocus do-rigne.	11000046
Ba már ocus ba binn in géim-sin, co mair 'géim uí Maigleine' ó shin ille oc drúthaib Éirenn.	11000047
Rogatad a chenn iar tain d'Fhergal, ocus rogatad a chenn don drúth.	11000048
Roboí mac alla géime in drúith 'sin áer co cenn trí lá ocus trí n-aidche.	11000049

cath_almaine_diplomatic

	badbh beldearg biorach, iolach imcenn f-Fergaile.
14000042	Scarais Murchadh ra midlaigh, brogais a triuna i-t-talmuin, do soí faobhar fria Fergal.
14000043	go f-fein dearmair des Almuin.
14000044	Bath ann cet ruirech rathach, cruadhach costadach carnach im naoí n-gelta gan míne. naoi mile fer n-armach.
14000045	Ceitri cet cabhsaidh a-Cruaich, las an amsaigh gaod san n-glíaidh latrí cedoibh Conaill cruaidh, asé.
14000046	Ra gabhadh annsain an druth ua Maighleine, & do radadh fair geim druith do dhénamh, & do rigne, bá már, & bá binn an gheim-sin, go mair geim ui Magleine ósin ale oc druthaibh Eirenn.
14000047	Ra gadadh achenn iarttain d-Fergal, & ra gadadh a cenn don

A graph for each re-use case (appears when you click on a point on the scatterplot)

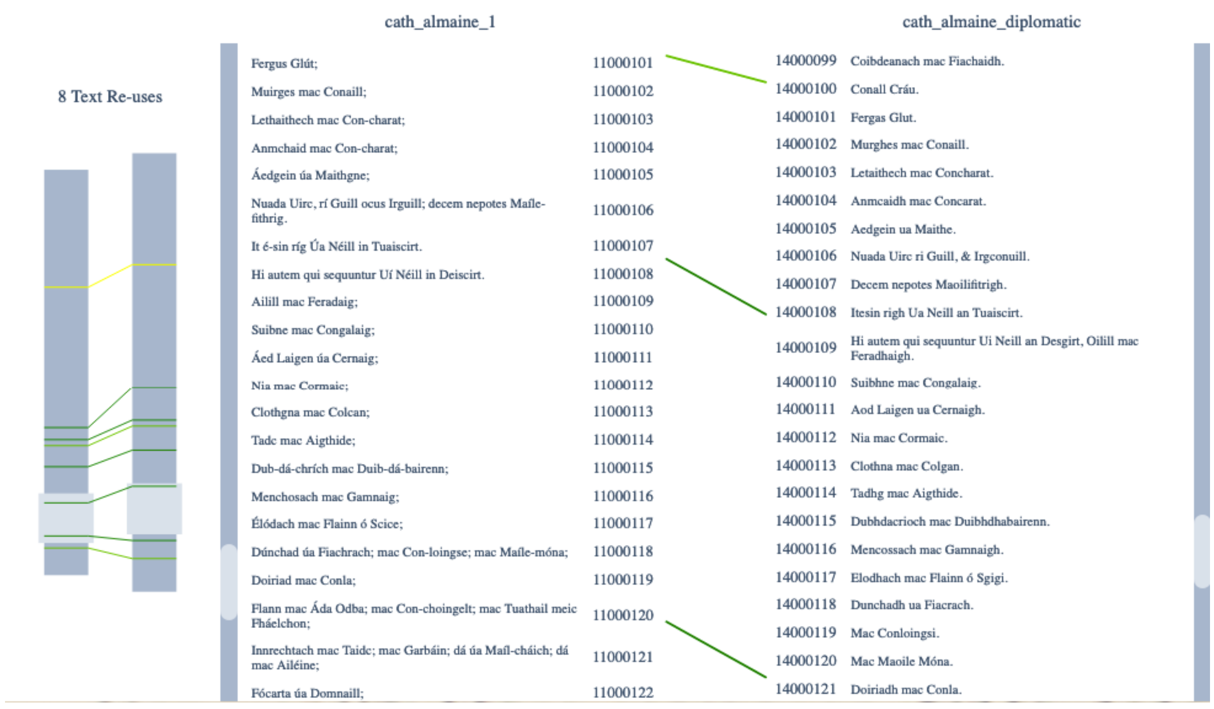


Text Re-use Alignment Visualization X

cath_almaine_1
Fergal mac Maile-dúin cum militibus suis;
cath_almaine_diplomatic
Fergal mac Maoilíduin cum militibus suis.

Stefan Jänicke, Leipzig University
DEV in BMBF-project *eTRACES* (PN: 01UA1101A)

Dynamic alignment

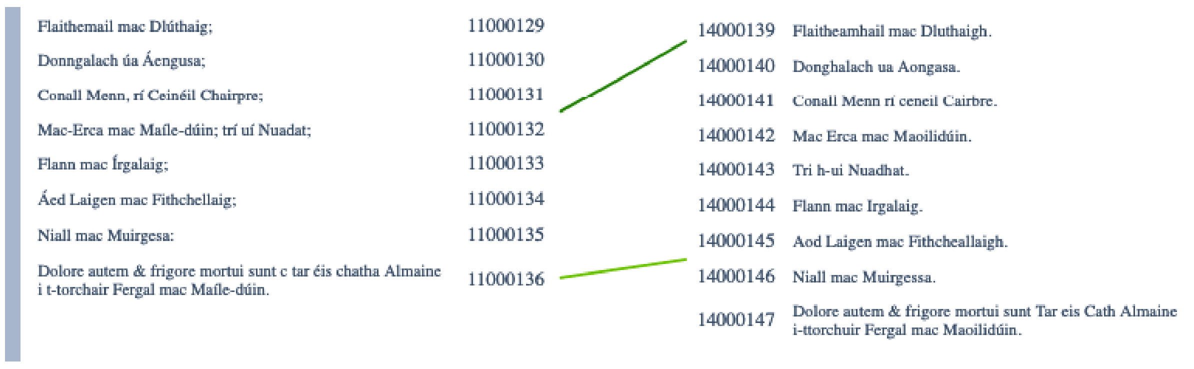


cath_almaine_1

Fergus Glút; 11000101
Muirges mac Conaill; 11000102
Lethaithech mac Con-charat; 11000103
Anmchaid mac Con-charat; 11000104
Áedgein úa Maithgne; 11000105
Nuada Uirc, rí Guill ocus Irguill; decem nepotes Maíle-fithrig. 11000106
It é-sin ríg Úa Néill in Tuaiscirt. 11000107
Hi autem qui sequuntur Uí Néill in Deiscirt. 11000108
Ailill mac Feradaig; 11000109
Suibne mac Congalaig; 11000110
Áed Laigen úa Cernaig; 11000111
Nia mac Cormaic; 11000112
Clothgna mac Colcan; 11000113
Tadc mac Aigthide; 11000114
Dub-dá-chrích mac Duib-dá-bairenn; 11000115
Menchosach mac Gamnaig; 11000116
Élódach mac Flainn ó Scice; 11000117
Dúnchad úa Fiachrach; mac Con-loingse; mac Maíle-móna; 11000118
Doiriad mac Conla; 11000119
Flann mac Áda Odba; mac Con-choingelt; mac Tuathail meic Fháelchon; 11000120
Innrechtach mac Taidc; mac Garbáin; dá úa Maíl-cháich; dá mac Ailéine; 11000121
Fócarta úa Domnaill; 11000122

cath_almaine_diplomatic

14000099 Coibdeanach mac Fiachaidh.
14000100 Conall Cráu.
14000101 Fergas Glut.
14000102 Murghes mac Conaill.
14000103 Letaithech mac Concharat.
14000104 Anmcaidh mac Concarat.
14000105 Aedgein ua Maithe.
14000106 Nuada Uirc ri Guill, & Irgconuill.
14000107 Decem nepotes Maoilifitrigh.
14000108 Itesin righ Ua Neill an Tuaiscirt.
14000109 Hi autem qui sequuntur Ui Neill an Desgirt, Oilill mac Feradhaigh.
14000110 Suibhne mac Congalaig.
14000111 Aod Laigen ua Cernaigh.
14000112 Nia mac Cormaic.
14000113 Clothna mac Colgan.
14000114 Tadhg mac Aigthide.
14000115 Dubhdacrioch mac Duibhdhabairenn.
14000116 Mencossach mac Gamnaigh.
14000117 Elodhach mac Flainn ó Sgigi.
14000118 Dunchadh ua Fiacrach.
14000119 Mac Conloingsi.
14000120 Mac Maoile Móna.
14000121 Doiriadh mac Conla.



Flaithemail mac Dlúthaig; 11000129
Donngalach úa Áengusa; 11000130
Conall Menn, rí Ceinéil Chairpre; 11000131
Mac-Erca mac Maíle-dúin; trí uí Nuadat; 11000132
Flann mac Írgalaig; 11000133
Áed Laigen mac Fithchellaig; 11000134
Niall mac Muirgesa; 11000135
Dolore autem & frigore mortui sunt c tar éis chatha Almaine i t-torchair Fergal mac Maíle-dúin. 11000136

14000139 Flaitheamhail mac Dluthaigh.
14000140 Donghalach ua Aongasa.
14000141 Conall Menn rí ceneil Cairbre.
14000142 Mac Erca mac Maoilidúin.
14000143 Tri h-ui Nuadhat.
14000144 Flann mac Irgalaig.
14000145 Aod Laigen mac Fithcheallaigh.
14000146 Niall mac Muirgessa.
14000147 Dolore autem & frigore mortui sunt Tar eis Cath Almaine i-ttorchuir Fergal mac Maoilidúin.

Paths

Names of the elements in red will be different from experiment to experiment, although structured in the same way.

To text reuse scores

tracer > data > corpora > your_corpus_name > TRACER_DATA > experiment_parameters > algorithm_chosen > second folder in a list > first folder in a list > threshold_chosen > **.score file**

To html with visualisation

tracer > data > corpora > your_corpus_name > TRACER_DATA > experiment_parameters > algorithm_chosen > second folder in a list > first folder in a list > threshold_chosen > -TraVizPosprocessingImpl- > html > **index.html**