Historical Text Reuse Detection Tutorial

Held at the *AIUCD Sixth Conference*, Rome, 23-24 January 2017

Organised by eTRAP in collaboration with DiXiT

Marco Büchler, Emily Franzini, Greta Franzini
[mbuechler|efranzini|gfranzini]@etrap.eu
Institute of Computer Science, University of Göttingen, Germany

Fabio Ciotti
fabio.ciotti@uniroma2.it
University of Roma Tor Vergata, Italy

The Organisers

eTRAP (**E**lectronic **T**ext **R**euse **A**cquisition **P**roject) is an Early Career Research Group funded by the German Federal Ministry of Education and Research (BMBF) and based at the University of Göttingen. The research group runs from 1st March 2015 until 28th February 2019. As the name suggests, this interdisciplinary team studies the linguistic and literary phenomenon that is *text reuse* with a particular focus on historical languages. More specifically, we look at how ancient authors copied, alluded to, paraphrased and translated each other as they spread their knowledge in writing. This early career research group seeks to provide a basic understanding of (historical) text reuse (it being distinct from plagiarism), and so to study what *defines* text reuse, *why* some people reuse information, *how* text is reused and how this practice has changed in history.
For more information about eTRAP and its projects please visit http://www.etrap.eu

DiXiT (Digital Scholarly Editions Initial Training Network) is an international network of high-profile institutions from the public and the private sector that are actively involved in the creation and publication of digital scholarly editions. DiXiT offers a coordinated training and research programme for early stage researchers and experienced researchers in the multi-disciplinary skills, technologies, theories, and methods of digital scholarly editing.
DiXiT is funded under Marie Curie Actions within the European Commission's 7th Framework Programme and runs from September 2013 until August 2017.

Tutorial Description

The Historical *Text Reuse Tutorial* builds on eTRAP's research activities, some of which deploy **TRACER** (http://www.etrap.eu/research/tracer/). TRACER is a suite of state-of-the-art Natural Language Processing (NLP) algorithms and functions aimed at discovering text reuse in multifarious corpora from multiple genres (in the same language). It is designed to work with historical text, such as Ancient Greek, Latin, Classical Arabic or medieval German, and provides researchers with a powerful engine to identify and display different types of text reuse ranging from verbatim quotations to paraphrase. To do so, TRACER implements basic NLP measures and operations and supplies an inbuilt step-wise pipeline which breaks down the challenging task of reuse detection into smaller sub-tasks. A human-readable and editable configuration file gives the user full control over the parameters during every step, thus accommodating specific needs.

TRACER can display reuse results through a text reuse alignment visualisation tool TRAViz (http://www.traviz.vizcovery.org/). Written in JavaScript, TRAViz retrieves the TRACER output and offers both close and distant views of the reuse for a more readable study of the texts.

Participants will learn how to run TRACER and how to display its results with TRAViz. In the afternoon of January 23 all participants will work on English data provided by the organisers in order to familiarise themselves with the software. During the morning session of January 24, and depending on the overall progress of the group, participants will be able to switch to their own datasets, provided these comply with the format required by TRACER, and will learn how to independently use TRACER.

Preliminary Schedule

The tutorial will run for **one and a half days** starting on the afternoon of Monday 23rd January and ending on the evening of Tuesday 24th.

Eligibility

Participants will have good knowledge of at least one operating system. Furthermore, participants are expected to be familiar with the principles of intertextuality studies and text reuse research.

Technical Requirements

A full list of tutorial requirements and instructions will be forwarded to the accepted participants, including package installations and expected data formatting. Accepted participants are asked to ensure that all requirements are fulfilled *before* the tutorial in order to avoid delaying the entire group on the day of the event.

Application and Bursaries

To apply for the tutorial, please send **a short CV** and a **brief motivation letter to contact@etrap.eu by Friday 16th December**. If successful, you must register for the AIUCD conference at: https://www.conftool.net/aiucd2017/

La Sapienza University makes available travel bursaries for early career researchers, who submit an abstract to the EADH day (for more information, see: http://aiucd2017.aiucd.it/?page_id=992). Should you be eligible for the bursary and wish to attend our tutorial, you must submit both an abstract to EADH and a CV with motivation letter to eTRAP. You may also apply for the tutorial without an EADH submission but you will not be eligible for a bursary in that case.

In order to provide everyone with adequate advice and assistance, the tutorial can accommodate **a maximum of 12 participants**. The tutorial is free of charge and will be conducted **in English**. However, any difficulties can also be discussed in Italian.

Previous Tutorials

This tutorial has been previously and successfully run in Göttingen (July 2015), London (September 2015), Tartu (October 2015), Galway (February 2016), Kraków (July 2016) and Venice (September 2016).