November 2015
Master's thesis project proposal

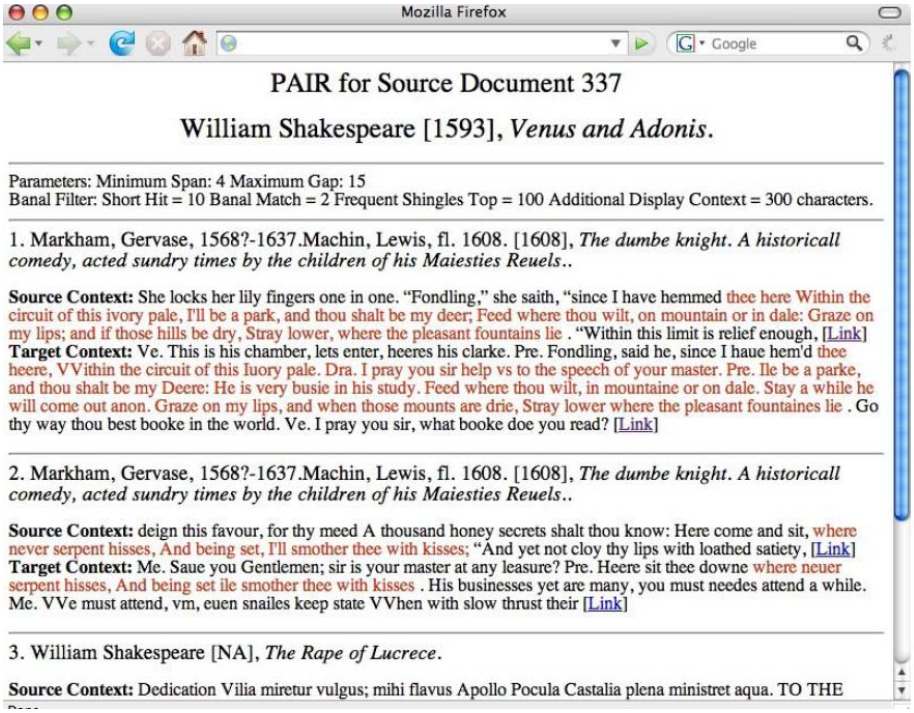
Proposal: Historical Text reuse (in Swedish Literature)

Keywords: sequence alignment; plagiarism; bitext alignment; intertextuality; historical text reuse; longest common substring(s); paraphrase; culturomics; evaluation

Goal: the goal of the Master thesis will be to apply (implement or adapt) techniques (e.g., borrowed from the field of bioinformatics) to identify lexically-similar passages (i.e. phrases, sentences, quotes, paraphrases) across collections of Swedish literary texts. Such techniques can use any suitable algorithms for that purpose, but preferably sequence alignment (Horton et al., 2010; Ganascia, 2011) and present/visualize the results in a user friendly (and navigable) way.

Background: according to Wikipedia <http://en.wikipedia.org/wiki/Sequence_alignment> a sequence alignment *in bioinformatics* "is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences". However, sequence alignments are also used for non-biological sequences, such as those present in natural language and also used for plagiarism detection (Clough, 2003).

Problem description: using the material in the Swedish Prose Fiction 1800-1900 database (SPF) and/or the Swedish Literature Bank (Litteraturbanken) you should implement or adapt and apply sequence alignment (and/or related algorithms) in order to detect common passages across the content of the collections (if they exist) and present them in a user friendly way, e.g. through web pages, or similar, with appropriate links of the original documents so that a literary scholar can easily navigate and further explore whether the algorithm has truly detected "text reuse" (e.g. long quotations) or perhaps **"**errors" produced by the process.



Example figure from the "Sequence alignment and the discovery of intertextual relations" by Mark Olsen.

As a special case you may instead choose to compare the written production of famous Swedish authors, such as August Strindberg's (born 1849) production, with all available material in SPF and Litteraturbanken (e.g., suitable material could be everything published from few years after August Strindberg's birth and back in time).

Evaluation

Text re-use is a broad and potentially vague area of research, and gold standards are scarce or not at all available (at least not for Swedish) in this area. Nevertheless, there are thinkable ways to perform a comprehensive evaluation for the results. For instance, in a comparable problem, in Information Retrieval, researchers have developed an evaluation method in terms of 11-points interpolated average precision (Manning et al., 2009:159) that can provide useful insights. Therefore, in this project you need also to consider and apply a method to evaluate the results you will acquire, in order to obtain an objective "big picture" view of the algorithm's performance.

Recommended skills: good programing skills in a script programming language, Python or Perl or and/or java (or other programing language you prefer) and (b) familiarity with XML and particularly the Text Encoding Initiative (TEI; <http://www.tei-c.org/index.xml>) structure (SPF and Litteraturbanken are encoded in TEI).

Note: The student that will take the assignment can either use or adapt existing software such as the "PAIR system" (Pairwise Alignment for Intertextual Relations) <http://code.google.com/p/text-pair/>; the MEDITE system <http://www-poleia.lip6.fr/~ganascia/Medite_Project>; the eTRACE (Büchler et al., 2014) or, even better, implement their own system.

References

- Büchler M., Burns P. R., Müller M., Franzini E. and Franzini G. (2014). Towards a Historical Text Re-use Detection. In *Text Mining, Theory and Applications of NLP*. Biemann, C. and Mehler, A. (eds). Pp. 221-238. Springer.
- Paul Clough. 2003. *Old and new challenges in automatic plagiarism detection*. University of Sheffield. Available from <http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism.pdf>
- Ganascia J-G. 2011. A Unilingual Text Aligner for Humanities. Application to Textual Genetics and to the Edition of Text Variants. Supporting Digital Humanities (SDH 2011), Copenhagen, Denmark. <http://www-poleia.lip6.fr/~ganascia/Medite_Project>
- Manning C.D., Raghavan P. and Schütze H. 2009. *An Introduction to Information Retrieval*. CUP Cambridge, UK <http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>
- Russell Horton, Mark Olsen and Glenn Roe. 2010. *Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections*. Digital Studies / Le champ numérique. Vol 2:1. Available from: <http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/190/235>
- Litteraturbanken: <http://litteraturbanken.se/#!/start>
- SPF: <http://spf1800-1900.se/#!/start>
- ...

Supervisors: Dimitrios Kokkinakis (Department of Swedish)**;** Mats Malm [domain expert] (Department of Literature, History of Ideas, and Religion)