A method for cleaning 19th century text with examples from *Transactions of the Royal Irish Academy 1800 - 1899*

Emma Clarke
ADAPT Centre, KDEG, Trinity College Dublin, Ireland
Clarkee8@tcd.ie

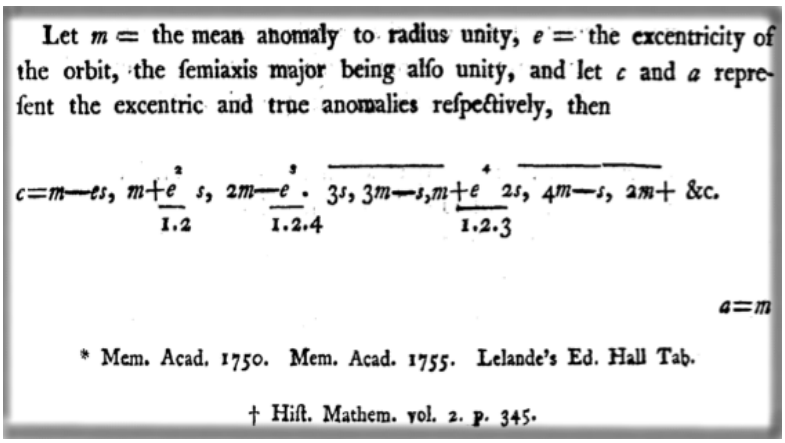
Alexander O'Connor
ADAPT Centre, KDEG, Trinity College Dublin, Ireland
Alex.Oconnor@scss.tcd.ie

This paper presents the cleaning pipeline, which was built on a corpus of nineteenth century scholarly articles - *Transactions of the Royal Irish Academy [RIA] 1800 – 1899*; and discusses the effect of different attempts to regularise and normalise the machine mediation of the insight.

One of the specific challenges which arises when attempting to carry out machine assisted textual analysis on a corpus such as Transactions is how to extract stable, insightful patterns from data which is difficult for the machine to understand. The articles within the corpus were all published between 1800 and 1899 and were digitised using OCR techniques. The digitisation was carried out as part of an unrelated project, a common situation for DH researchers, which limits control over quality and suitability.

This corpus presented characteristic challenges incorporated in the articles. The following factors had to be taken into consideration when cleaning the *Transactions* corpus (see figure 1 for specific examples from Transactions of the RIA (Brinkley, 1803)):

- the inclusion of technical and structural content (tables, references, symbols, equations)
- the use of multilingual and specialised language
- the importance of multi-word expressions (Sag et al., 2002)
- polysemy (words or phrases which have many different meanings)
- homonymy (words which are spelled the same, but have different meanings)
- other linguistic variations (abbreviations: Vol. for volume, Fig. for figure etc.)



Let m = the mean anomaly to radius unity, e = the excentricity of the orbit, the ſemiaxis major being alſo unity, and let c and a repreſent the excentric and true anomalies reſpectively, then

$$c = m - es,\ m + \frac{e^2}{1.2}s,\ 2m - \frac{e^3}{1.2.4}\cdot 3s,\ 3m - s,\ m + \frac{e^4}{1.2.3}2s,\ 4m - s,\ 2m + \text{ &c.}$$

$a = m$

* Mem. Acad. 1750. Mem. Acad. 1755. Lelande's Ed. Hall Tab.

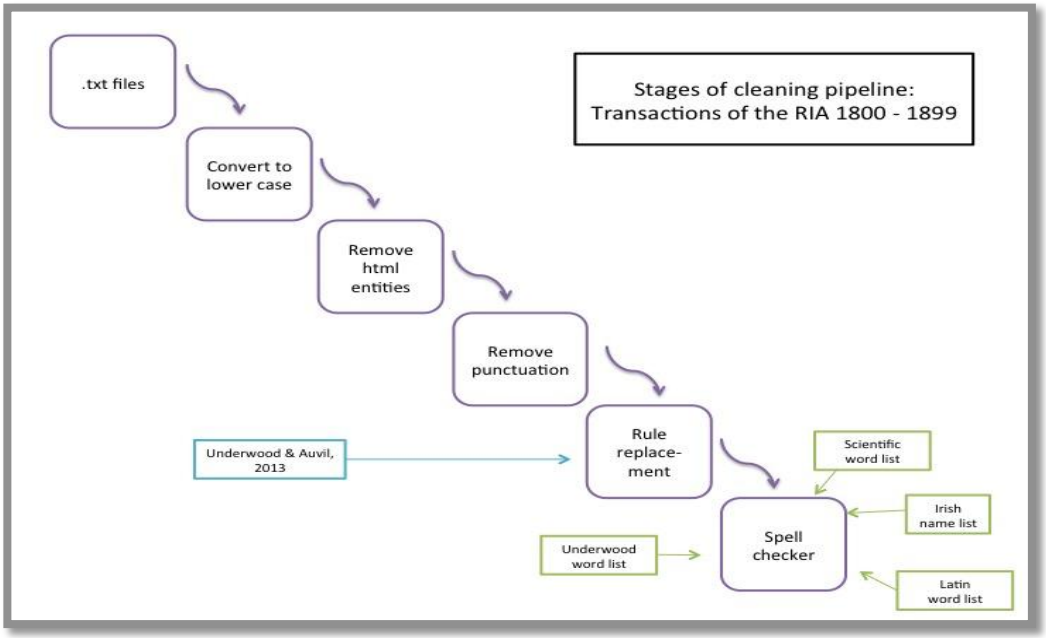
† Hiſt. Mathem. vol. 2. p. 345.

Figure 1: examples of challenging factors in Transactions of RIA

Cleaning process and collection of resources

One of the key challenges in any automatic approach is the need to address the inconsistency, complexity and general noise which exists in real data. This is especially true in the case of a corpus such as this, as the content is over a wide diachronic range (affecting the typography amongst other things), has high subject-matter heterogeneity, and was converted using OCR techniques. A key aspect of this work was that there was no access to the underlying documents, to higher-quality OCR, or to reference texts. This meant that many of the approaches to performance improvement were not available.

Previous approaches to cleaning 19th century OCR text include Underwood's normalisation of OCR errors (Underwood, 2013) and Jockers' combination of an expanded stop word list and part of speech tagging (Jockers, 2013). However, due to the mixed & technical nature of the corpus content and the fact that we needed to remove structural content such as tables, charts and figures, we chose the approach outlined in this flowchart (Underwood and Auvil, n.d.); (Underwood, 2012); (Petrie, n.d.):



Python scripts were developed to carry out each stage of the cleaning pipeline and these are available in the following github repository: https://github.com/emmaclarke/TransactionsRIA_1800-99.

Conclusion

The paper provides insight into an increasingly common challenge facing digital researchers: the artifacts which they wish to investigate are complex, incorporating significant variation in language, formatting, typesetting and structure. Moreover, the artifacts under investigation are the 'born-digital' incarnations: no access is available to the originals, and the collection is small and unusual. The approach that we present is therefore relevant for many studies which seek to gain insight into the implicit meaning in document corpora, where complete control is not possible. The positive improvements found with this approach, along with discussions on how to customise, and more formally evaluate, for similar work in future, are intended to show the underlying value of the work beyond the conclusions drawn about these texts.

References

Brinkley, J., 1803. An Examination of Various Solutions of Kepler's Problem, and a Short practical Solution of That Problem Pointed out. Trans. R. Ir. Acad. 83–131.

Jockers, M., 2013. "Secret" Recipe for Topic Modeling Themes [WWW Document]. Matthew Jockers. URL http://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/ (accessed 8.26.13).

Petrie, J., n.d. Scientific word list for spell-checkers/spelling dictionaries [WWW Document]. John Petrie's LifeBlag. URL http://www.jpetrie.net/scientific-word-list-for-spell-checkersspelling-dictionaries/ (accessed 8.18.13).

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., 2002. Multiword Expressions: A Pain in the Neck for NLP, in: Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1–15.

Underwood, T., Auvil, L., n.d. Basic OCR correction [WWW Document]. Uses Scale Lit. Study. URL http://usesofscale.com/gritty-details/basic-ocr-correction/ (accessed 6.25.13).

Underwood, Ted, 2013. A half-decent OCR normalizer for English texts after 1700. Stone Shell. URL http://tedunderwood.com/2013/12/10/a-half-decent-ocr-normalizer-for-english-texts-after-1700/ (accessed 6.25.13)

Underwood, Ted, 2012. Flowchart for probabilistic OCR correction. Uses Scale Lit. Study. URL http://usesofscale.com/2012/10/14/probabilisticocr/ (accessed 6.25.13)