Mapping *Finnegans Wake* scholarship: "the elucidation of complications" (*FW* 109.05)

Richard Barlow and Jonathan Reeve

This report details progress being made on a project titled 'Mapping *Finnegans Wake* scholarship: Creating an online research platform linking the full text of *Finnegans Wake* text to existing analysis.' The project began in 2018 and involves academics based at Nanyang Technological University, Columbia University, and the Centre for Manuscript Genetics, University of Antwerp. The aim of the project is to link scholarship on James Joyce's *Finnegans Wake* (1939) from the digital library JSTOR to a digital text of the book. Joyce (1882-1941) is a hugely important writer with a massive scholarly industry dedicated to his works. *Finnegans Wake* is his final text and is a central text of literary Modernism. However, due to the difficulties and complexity of the text (it is an extremely rich text containing lexical elements from roughly seventy different languages, numerous complicated plots and 'nodal structures', and extensive use of allusion and multilingual wordplay) it remains a great challenge for the reader and is often described as 'unreadable.'[1]

As writers such as Jacques Derrida and Steven Connor have suggested, *Finnegans Wake* is something like a premonition of the internet or the information age, in that it stores so much interrelated information and that it behaves like a hypertext. Furthermore, its 'nodal systems' invite readings that connect disparate points within the text while its references and allusions are similar to the way data is connected online.[2] Computers and the internet—especially annotation sites like fweet.org—have changed the way we read *Finnegans Wake*. We are less likely to experience the text in a fully linear and 'physical' way these days—at least in academia—and more likely to seek out complex textual systems and patterns through the use of information technology. Despite the affinities between *Finnegans Wake* and information technology, there is currently no website linking the text of *Finnegans Wake* to scholarly articles about its various sections. In other words, there is nothing digitally linking the primary text to the secondary texts. At the moment, there are only non-university aligned websites providing elucidations/annotations to the text (such as fweet.org). Furthermore, there is currently no 'map' to tell us which sections of the text have been researched and which sections remain critical 'blind spots.' Our digital platform will contain a digital 'map' detailing which sections of the text of *Finnegans Wake* have received high levels of attention/scholarship and which have remained relatively

unexamined.

The completed platform will include an online text of *Finnegans Wake* (in public domain since 2013) with hyperlinks to secondary material in JSTOR, initially on work available in the flagship Joyce studies journal *James Joyce Quarterly*. Users studying a specific section of *Finnegans Wake* will be able to see at-a-glance if any scholarship on that section is available. If the user has access to JSTOR (through their university for example), they will be able to access the material via hyperlink. As such, the proposed research platform will be of interest and use to the Joyce studies/Irish studies/Modernism studies communities both academic and non-academic. The research platform will be especially useful for scholars working on words/sentences/sections of *Finnegans Wake* as they will be able to instantly ascertain what work has already been carried out. Our research platform might also encourage non-academics to read and work with the text.

After an preliminary presentation at the 'Finnegans Wake at 80' conference held at Trinity College Dublin in April 2019, the project sourced a digital text of *Finnegans Wake* (courtesy of Dr Wim Van Mierlo, University of Loughborough). In January 2020 the project received a large customized dataset from JSTOR via their 'Data for Research' program. The dataset consisted of all the journal articles on Joyce's *Finnegans Wake*. Following work by Jonathan Reeve, we have found that there are three main *Finnegans Wake* critical hotspots (i.e. three areas in the novel most frequented quoted in the academic literature): these are the very 'beginning' of the text, the very 'end' of the text, and this very long sentence from about 28% of the way through the book:

> The warped flooring of the lair and soundconducting walls thereof, to say nothing of the uprights and imposts, were persianly literatured with burst loveletters, telltale stories, stickyback snaps, doubtful eggshells, bouchers, flints, borers, puffers, amygdaloid almonds, rindless raisins, alphybettyformed verbage, vivlical viasses, ompiter dictas, visus umbique, ahems and ahahs, ineffible tries at speech unasyllabled, you owe mes, eyoldhyms, fluefoul smut, fallen lucifers, vestas which had served, showered ornaments, borrowed brogues, reversibles jackets, blackeye lenses, family jars, falsehair shirts, Godforsaken scapulars, neverworn breeches, cutthroat ties, counterfeit franks, best intentions, curried notes, upset latten tintacks, unused mill and stumpling stones, twisted quills, painful digests, magnifying wineglasses, solid objects cast at goblins, once current puns, quashed quotatoes, messes of mottage, unquestionable issue papers, seedy ejaculations, limerick damns, crocodile tears, spilt ink, blasphematory spits, stale shestnuts, schoolgirl's, young ladies', milkmaids', washerwomen's, shopkeepers' wives, merry widows', ex nuns', vice abbess's, pro virgins', super whores', silent sisters', Charleys' aunts', grandmothers', mothers'-in-laws, fostermothers', godmothers' garters, tress clippings from right, lift and cintrum, worms of snot, toothsome pickings, cans of Swiss condensed bilk, highbrow lotions, kisses from

> the antipodes, presents from pickpockets, borrowed plumes, relaxable handgrips, princess promises, lees of whine, deoxodised carbons, convertible collars, diviliouker doffers, broken wafers, unloosed shoe latchets, crooked strait waistcoats, fresh horrors from Hades, globules of mercury, undeleted glete, glass eyes for an eye, gloss teeth for a tooth, war moans, special sighs, longsufferings of longstanding, ahs ohs ouis sis jas jos gias neys thaws sos, yeses and yeses and yeses, to which, if one has the stomach to add the breakages, upheavals distortions, inversions of all this chambermade music one stands, given a grain of goodwill, a fair chance of actually seeing the whirling dervish, Tumult, son of Thunder, self exiled in upon his ego, a nightlong a shaking betwixtween white or reddr hawrors, noondayterrorised to skin and bone by an ineluctable phantom (may the Shaper have mercery on him!) writing the mystery of himsel in furniture. (*FW* 183.8– 184.10)

This section of *Finnegans Wake* shows the artist figure Shem at work (Shem is based on Joyce himself). Shem's method of composition – the accumulation of random, everyday bric-a-brac – is also a self-reflexive commentary on Joyce' method of composing *Finnegans Wake*. Along with a structure based on Giambattista Vico's *The New Science* (1725) and the use of multiple languages and wordplay, the gathering and adapting arrangement of pre-existing or 'found' literary, cultural, or historical materials (including squashed quotations or "quashed quotatoes") was a central artistic technique in Joyce's creation of the text (perhaps reflecting the confusing mergings and transformations of dreams). As such, it is easy to see why this section has become a critical hotspot – it can be used to discuss or explain the Modernist bricolage of *Finnegans Wake* itself. Meanwhile, the 'beginning' and 'ending' of the text have frequently been used to demonstrate the cyclical nature of the text – that it has no real 'beginning' or 'ending'. As is well known, *Finnegans Wake* 'begins' (or begins again, depending) in the middle of a sentence "riverrun, past Eve and Adam's, from swerve of shore to bend of bay, brings us by a commodius vicus of recirculation back to Howth Castle and Environs" (*FW* 3.1–3). The 'beginning' of that sentence can be found at the 'end' of the book: "Finn, again! Take. Bussoftlhee, mememormee! Till thousendsthee. Lps. The keys to. Given! A way a lone a last a loved a long the" (*FW* 628.14–16) Thus the text's structure replicates a central theme of the text itself – repetitions and returns (demonstrated in its preoccupations with cycles of history, literary recycling, and different forms of resurrection).

To find these areas of critical interest, we use the text reuse detection program *Text-matcher*, initially written by Jonathan Reeve for the Middlemarch Critical Histories project (Reeve et al., 2017). Text reuse detection, first developed for commercial uses such as plagiarism detection, is beginning to be used as an analytic tool in literary studies (Piper and Manalad, 2020). Our algorithm operates in two passes. The first compares lemma trigrams using Python difflib's SequenceMatcher, which as the module's authors describe it, 'predates, and is a little fancier than, an algorithm published in the late 1980's by Ratcliff

and Obershelp under the hyperbolic name "gestalt pattern matching" ' (Peters, 2016). This library matches text approximately, automatically ignoring textual differences it considers 'junk', or differences that would be unimportant to most human readers. This allows for a first-pass fuzziness. From there, *Text-matcher* expands the match in either direction, comparing Levenshtein edit distances between candidate lemmas, and ignoring punctuation, line breaks, and paratext such as page numbers and footnotes. This allows us to avoid many of the difficulties that arise from comparing text that contain errors from their optical character recognition. Crucially, it also ignores XML tags, which enables us to run this program over a TEI XML source text: an edition of *Finnegans Wake* encoded in the eXtensible Markup Language of the Text Encoding Initiative.[3]

From the beginning, we didn't want this project to be restricted to just the *Wake*, but to build a repeatable, standards-focused framework which could be used in other applications. Toward this end, we chose to encode data in TEI XML, such that our digital edition, accompanying annotations, and all related information would be available for future projects to remix and reuse. This follows the methodology used by Open Editions, a project to which this experiment contributes.[4]

However, since XML is notorious for certain limitations, such as its inability to incorporate overlapping tags, we use a relatively new TEI feature, the `<standOff>` tag, which allows us to maintain a separate file containing links from passages in our source XML to the JSTOR articles that quote them, with character offsets and bibliographic metadata for each. These two files—the text and annotations—are then transformed into HTML using a custom script written in the Haskell programming language, and served as a static website. The result is a interactive edition, similar to that of JSTOR Labs's own *Understanding* series, in which textual passages link to critical articles that discuss them.[5]

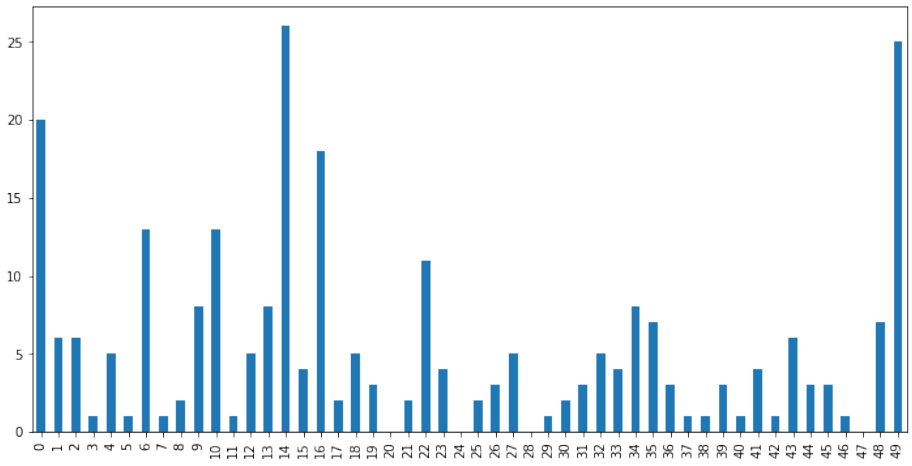

Figure 1: Quotations in Narrative Time

The more important insights of this project, however, come not from the creation of a product like a website, but from the data generated from an analysis of its content. Fig. 1 shows the number of quotations of the *Wake* found in the secondary literature, according to their position in the novel, with the novel's first passages on the left, and the last on the right. With one exception, the beginning and the end are quoted the most, and the the third quarter of each novel is the least quoted.

Our hypotheses for this phenomenon are varied. First, it could be that this area is just where the critic's interest flags: with any long work, there is an associated reading fatigue. This is seemingly confirmed by the similarities between our *Wake* quotation distribution, and those found of *Middlemarch* (Reeve et al., 2017). The Stanford Literary Lab finds a similar structure of quotation distribution in narrative time, by comparing a large corpus with articles from British Periodicals Online.

Another hypothesis is that there is a correlation between the novel's intelligibility and its quotability. To test this, we model intelligibility using three metrics. First, we compute the proportion of words which the spell checker Hunspell identifies as misspelled, given a dictionary of British English. Next, we find the proportion of sentences which are correctly identified as English, by the Python Langdetect library, an adaptation of an algorithm originally developed at Google by Nakatani Shuyo (Danilák, 2018). Finally, we compute the Coleman-Liau readability index, CLI, given by the following equation (Coleman and Liau, 1975).

$$CLI = 0.0588(\frac{\#letters}{\#words}) - 0.296(\frac{\#sentences}{\#words}) - 15.8$$

These scores, normalized, and computed across the same 50 novel segments, is shown in Fig. 2.

There are a few notable trends apparent in this model of the novel's intelligibility. First, these scores don't all seem to agree. While the proportion of English-like sentences seems to follow the Coleman-Liau index, the lowest proportion of nonstandard spelling is in segment 26, which corresponds with the highest CLI. However, this does partially confirm some of our suspicions: segments 23 and 24 have the lowest CLI values, and proportions of sentences inferred as English, and these are among the least-cited segments. In fact, segment 24 has no citations at all. Similarly, segments 38 and 39, which score low in these two metrics, have equally low numbers of quotations in the secondary literature.

If we examine the date ranges of these quotations, as shown in Fig. 1, we see that the greatest number of them come after 1986. While some of this trend may be attributable to an uneven availability of journals digitized for JSTOR, it is nonetheless suggestive of a significant jump in critical attention to the *Wake* around this time, possibly influenced by its publication trend: the greatest number of *Wake* editions, in the history of its publication, appear only a few years

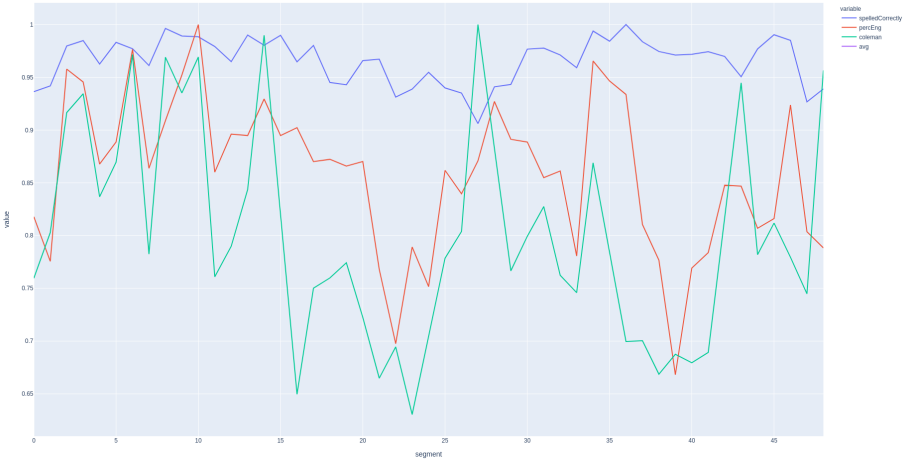

Figure 2: Intelligibility Scores by Novel Segment

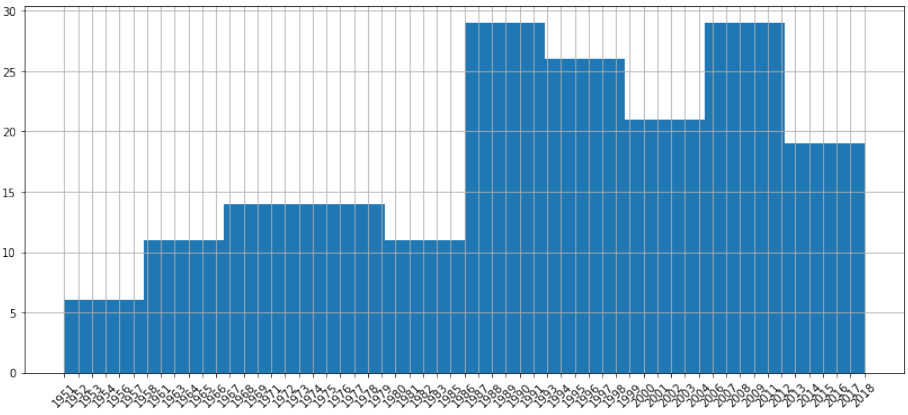

Figure 3: Quotations by Year of Quotation

before, in the late '70s and early '80s, according to our analysis of publication data from the Open Library API.[6] This trend coincides with the appearance of the term *Finnegans Wake* in the Google and Hathi Trust datasets, as provided by the Ngrams Viewer service.[7]

All of the data and code used to create this project is freely available on GitHub, via Open Editions, and is licensed under the GNU Public License, Version 3.[8] We encourage others to reproduce these experiments, and create their own analyses using our data. Our next steps are to produce similar critically-annotated editions of the remainder of Joyce's major works, and to further generalize this framework, so that it may be more widely used.

Funding

This work was supported by the Singapore Ministry of Education's Tier 1 Academic Research Fund.

Notes

[1] See page vii of Seamus Deane's introduction to the Penguin edition of *Finnegans Wake* (London, 1992), for example.

[2] "The *Wake* seems to model itself, not on the newspaper, as *Ulysses* seemed to do, but on the culture of electronic communications which was inaugurated in 1876 with the near-simultaneous invention of the telephone and the phonograph and accelerated in the early decades of the twentieth century with the rapid development of radio, cinema, and, from the mid-1920s, television ... *Finnegans Wake* may be said to predict and exemplify the age of electronic media. Electronic media are the fulfilment of the scientific promise of universal convertibility of forces ... It is perhaps not surprising then that the increasing interest in applying contemporary computer technology to the study and reading of Joyce should begin to disclose a profound affinity between such technologies and their object. If *Ulysses* and *Finnegans Wake* call for the resources of hypertext and multimedia databases to make visible and available the wealth of interconnections of which each consists, then this is perhaps partly because the works themselves appear singly or collectively to be what Derrida, again spurred into Wakean imitation, has called a 'programotelephonic encyclopaedia' (Connor, 2018). See also: "this 1000th generation computer – *Ulysses*, *Finnegans Wake* – besides which the current technology of our computers and our micro-computerified archives and our translating machines remains a bricolage of a prehistoric child's toys. And above all its mechanisms are of a slowness incommensurable with the quasi-infinite speed of the movements on Joyce's cables. How could you calculate the speed with which a mark, a marked piece of information, is placed in contact with another in the same word or from one end of the book to another?" Derrida also discusses 'the double or the simulation of the event ''Joyce'', the name of Joyce, the signed work, the Joyce software today, joyceware' (Attridge, 2017).

[3] For an introduction to the TEI, see Cummings (2013).

[4] Open Editions, at open-editions.org, is first described in Reeve and Gabler (2019), and encompasses a specification for the creation of richly-annotated TEI XML scholarly editions, along with a software stack that manages and publishes them.

[5] See https://www.jstor.org/understand/

[6]For more details on the publication history of *Finnegans Wake*, see our analysis notebooks, at, e.g., https://github.com/open-editions/corpus-joyce-finnegans-wake-tei/tree/master/criticism-analysis/metadata-analysis.ipynb.

[7]For more details, search for the term *Finnegans Wake* in https://books.google.com/ngrams/

[8]See https://github.com/open-editions/corpus-joyce-finnegans-wake-tei

References

Attridge, D. (2017) 'Ulysses Gramophone: Hear Say Yes in Joyce', in *Acts of Literature*. New York: Routledge. pp. 253–309.

Coleman, M. & Liau, T. L. (1975) A computer readability formula designed for machine scoring. *Journal of Applied Psychology*. 60 (2), 283.

Connor, S. (2018) *James Joyce*. Liverpool: Liverpool University Press.

Cummings, J. (2013) 'The Text Encoding Initiative and the Study of Literature', in *A Companion to Digital Literary Studies*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/9781405177504.ch25.

Danilák, M. (2018) *Langdetect*. https://github.com/Mimino666/langdetect (Accessed 25 December 2020).

Joyce, J. (1939) *Finnegans Wake*. Viking Press.

Peters, T. (2016) *Difflib*. Python Software Foundation. https://docs.python.org/3.5/_sources/library/difflib.txt (Accessed 19 March 2016).

Piper, A. & Manalad, J. (2020) Measuring unreading. *Goethe Yearbook*. 27233–241.

Reeve, J. et al. (2017) 'Frequently cited passages across time: New methods for studying the critical reception of texts.', in *Digital Humanities 2017 Book of Abstracts*. https://dh2017.adho.org/abstracts/264/264.pdf.

Reeve, J. & Gabler, H. W. (2019) Open Editions Online. *James Joyce Quarterly*. 57 (1), 163–172.